Étude et développement d'un système d'estimation tridimensionnelle de trajectoire de corps humain en mouvement

Yann Desmarais

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'INSTITUT MINES-TÉLÉCOM (IMT) -
ÉCOLE NATIONALE SUPÉRIEURE DES MINES D'ALÈS (IMT MINES ALÈS)

En Informatique

École doctorale : I2S
Portée par l'Université de Montpellier

Unité de recherche : Euromov Digital Health in Motion

Étude et développement d'un système d'estimation tridimensionnelle de trajectoire de corps humain en mouvement

Présentée par Desmarais Yann
Le 13 Juin 2023

Sous la direction de Philippe Montesinos
et Denis Mottet
et Pierre Slangen

Devant le jury composé de :

Jenny BENOIS-PINEAU, PR, LABRI, Université de Bordeaux	Rapportrice
Mohamed DAOUDI, PR, CRIStAL, IMT Nord Europe	Rapporteur
Valérie GOUET-BRUNET, DR, LASTIG, University Gustave Eiffel	Examinatrice, Présidente
Olivier STRAUSS, MCF, LIRMM, Université de Montpellier	Examinateur
Phillipe MONTESINOS, M.A., EuroMov DHM, IMT Mines Alès	Directeur
Denis MOTTET, PR, EuroMov DHM, Université de Montpellier	Codirecteur
Pierre SLANGEN, PR, EuroMov DHM, IMT Mines Alès	Codirecteur

COLLÈGE DOCTORAL UNIVERSITÉ DE MONTPELLIER

IMT Mines Alès
École Mines-Télécom

[...] ; de même que nous ne percevons jamais un mouvement que comme une série de points isolés, en réalité nous ne le voyons donc pas, nous y inférons. La soudaineté que mettent certains effets à se détacher nous induit en erreur ; cependant cette soudaineté n'existe que pour nous. Dans cette seconde de soudaineté il y a une infinité de phénomènes qui nous échappent.

Le Gai Savoir
FRIEDRICH NIETZSCHE

À mes grand-pères Pierre Vigneron et Gaston Desmarais

Remerciements

Je tiens d'abord à remercier mes directeurs de thèse qui m'ont fait confiance et qui m'ont grandement apporté au cours de cette thèse, chacun dans leur domaine d'expertise : Philippe au laboratoire et toujours prêt à prendre sur son temps pour m'expliquer un point important, Denis à Euromov qui m'a fait découvrir les sciences du mouvement humain et Pierre pour l'assistance précieuse dans les différentes plateformes de capture du mouvement à Alès. Vous m'avez apporté un concours précieux tout au long de ces années et j'ai beaucoup apprécié travailler avec vous.

Je remercie aussi Euromov Digital Health in Motion et sa direction, Jacky Montmain et Stéphane Perrey qui ont permis la préparation de cette thèse au sein de l'unité. Je tiens aussi à remercier également laboratoire du CERIS et l'équipe I3A et l'ensemble des chercheurs qui les composent qui m'ont accueilli. Je remercie tout particulièrement le laboratoire à Alès qui m'a toujours apporté le soutien logistique et matériel nécessaire au développement du prototype, et ce, malgré les années difficiles dues à la pandémie de Covid-19.

Ce soutien était également apporté par le personnel administratif et technique du laboratoire, tout particulièrement Edith Teychene qui est venu à notre secours de nombreuses fois pour les soucis administratifs. Je remercie aussi Pierre Jean, Pierre-Antoine Jean pour leur assistance technique (notamment pour le serveur de calcul) et leur gentillesse. Je remercie aussi Pierre Richard qui fût d'une aide cruciale pour résoudre les problèmes matériels et systèmes liés à l'installation des différentes machines que j'ai pu utiliser.

Je tiens aussi à remercier tous les doctorants et stagiaires, collègues de bureau ou non qui ont animé mes journées pendant ces trois années au laboratoire. Tout d'abord l'équipe de babyfoot du midi : Sakhi, Yu, Nassir et Bastien et puis tous les autres : Jihane, Quentin, Ali, Hamza, Lucie, Cédric, Alexandre et Jérémy et bien d'autres.

Pour finir, je veux adresser un dernier grand merci à toute ma famille qui m'a soutenu, y compris pendant les moments de doutes pendant tout ce long cheminement qu'est la thèse. Merci à mes parents Suzanne et Erick, à mes frères Cedric et Damien et mes grands-mères Jacqueline et Micheline et à toutes mes tantes, oncles et cousins. Et surtout merci à toi Maxime que j'ai rencontré au cours de ce périple et sans qui je ne sais pas si je serais allé au bout.

Résumé

L'étude du mouvement humain requiert l'estimation de la posture la plus précise possible. Ce sont souvent des systèmes de capture du mouvement optique qui sont employés, nécessitant le placement de marqueurs sur le sujet étudié. Ces marqueurs sont ensuite détectés dans la scène par plusieurs caméras infrarouges, ce qui permet de les reconstruire en trois dimensions et de les suivre au cours du mouvement filmé. Ce dispositif permet ainsi de déduire une posture juste même si l'environnement dans lequel le sujet évolue est partiellement occulté par des objets. Cependant, ces systèmes sont très coûteux et demandent la plupart du temps une installation fixe en laboratoire qui permet peu de flexibilité. De plus, le placement des marqueurs peut s'avérer trop intrusif et peut altérer les mouvements dans le cadre de certaines études.

Dans cette thèse, nous étudions la possibilité d'un dispositif sans marqueurs, robuste, précis, peu coûteux et permettant la capture du mouvement sur une zone de la taille d'une pièce ou d'une habitation. Ce système sera composé de caméras à bas coût, afin de pouvoir multiplier les angles de vues et limiter les risques d'occultation.

Récemment, des méthodes de vision par ordinateur permettant l'extraction de la posture directement dans des images sont devenues de plus en plus exactes. Elles utilisent l'apprentissage profond grâce à des jeux de donnée à grande à échelle labellisés avec la capture de mouvement optique. Ces solutions ouvrent la voie à la capture de mouvement sans marqueurs. C'est dans ce cadre que s'inscrit la méthode utilisée par le prototype développé. La problématique étudiée consiste à combiner l'estimation de posture avec un modèle d'apprentissage machine et la reconstruction en trois dimensions utilisant la géométrie projective pour exploiter les multiples angles de vue capturés.

Le prototype final est constitué de quatre micro-ordinateurs équipés de caméras permettant une acquisition simultanée. Les caméras sont calibrées, ce qui permet une reconstruction robuste de la posture en trois dimensions à partir de l'inférence des postures en deux dimensions dans les images de chaque vue.

Le modèle d'estimation de posture utilisé par le prototype est évalué sur un jeu de donnée de référence et sa précision est équivalente aux autres méthodes de l'état de l'art. De plus, le prototype complet est testé en laboratoire parallèlement à un système de capture de mouvement optique. Cette expérience permet de valider le système et les différentes étapes de calibration, acquisition puis reconstruction de la posture sur plusieurs sujets et mouvements.

Table des matières

Table des figures

Chapitre 1

Introduction

1.1 Historique

L'étude scientifique du mouvement animal et humain est, depuis ses débuts au 19e siècle, étroitement liée aux progrès techniques des méthodes d'acquisitions. Les expériences avec de multiples chambres photographiques sur des pistes hippiques d'E. Muybridge ou l'invention de la chronophotographie par Marey (1882) en se basant sur le "revolver photographique" le montre. Les dispositifs précurseurs de capture scientifique du mouvement humain sont sans doute attribuables à E-J. Marey. Ses "chaussures exploratrices" permettent la mesure directe du rythme de marche tandis que sa "combinaison de chronophotographie" préfigure les combinaisons avec marqueurs photo-réfléchissants des systèmes de capture du mouvement contemporains (Marey, 1891). C'est avec l'informatique moderne et les avancées de la recherche en vision par ordinateur, en reconstruction 3D notamment, que les premiers systèmes optiques de capture du mouvement ont vu le jour.

Le corps humain en mouvement est représenté au cinéma dès sa création, mais c'est dans le courant des années 90 que se généralise l'utilisation de la capture de mouvement pour produire des personnages en image de synthèse. Les studios de cinéma utilisent des systèmes complexes dotés de plus en plus de caméras qui identifient et suivent des marqueurs réfléchissant sur des combinaisons que revêtent les acteurs. Le monde du jeu vidéo fait également l'utilisation de cette technique qui permet d'obtenir des séquences avec des avatars numériques articulés qui peuvent aisément être animés.

1.2 Contexte général

De nombreux champs d'application font usage de la posture humaine ou de sa sémantique, que ce soit dans le divertissement (cinéma d'animation, jeux vidéos...) ou dans un cadre scientifique (analyse du mouvement pathologique ou sportif). Avec internet et les médias sociaux, la recrudescence d'images de personnes et les performances en vision par ordinateur des méthodes d'apprentissage profond ont motivé les grandes entreprises du numérique à mettre au point leurs propres méthodes de reconnaissance dans de nombreux domaines, dont l'estimation de posture humaine.

C'est dans ce contexte que l'exactitude des méthodes de l'état de l'art est régulièrement dépassée depuis le milieu des années 2010 pour les tâches d'estimation de posture 2D puis, plus récemment, pour la 3D. La motivation pour le développement de ces méthodes est, d'une part pour la mise en place de meilleures interactions homme-machine pour les nouveaux périphériques et interfaces, mais aussi et surtout un besoin croissant de monitoring et diagnostic des troubles musculo-squelettiques qui va de pair avec le vieillissement croissant de la population. Aujourd'hui, la capture de mouvement est de plus en plus accessible. En 2010, Microsoft met en vente le capteur Kinect qui permet de capturer la pose humaine à partir d'images RVB et de profondeur. Cependant, il est principalement destiné à un usage vidéoludique et ne convient pas à l'acquisition en extérieur. Plus récemment, de nombreuses solutions commerciales dites "sans-marqueurs" ont vu le jour et proposent de reproduire des résultats similaires avec seulement des images ou vidéos. La nécessité de placer des marqueurs sur les sujets suivis tend à disparaître.

Cependant, les algorithmes de détection de personnes et de posture basés uniquement sur les caractéristiques visuelles existent depuis longtemps (O'Rourke and Badler (1979), O'Rourke and Badler (1980)). Historiquement, ces algorithmes adaptaient des caractéristiques prédéfinies à des modèles humains complexes par parties basés sur les membres (représentations de la silhouette avec des cylindres, des figures-bâtons, des mailles, des cônes ou des boîtes (Fischler and Elschlager (1973), Felzenszwalb and Huttenlocher (2005)). Bien que certaines techniques modernes utilisent encore cette approche, la partie visant à extraire les caractéristiques des images est désormais quasi exclusivement réalisée à l'aide de réseaux de neurones convolutifs.

L'extraction des mouvements est de plus en plus accessible et de moins en moins contrainte. Les tout premiers systèmes historiques étaient mécaniques et demandaient du matériel spécialisé, complexe. Les méthodes modernes de capture du mouvement sont des systèmes optiques nécessitant des combinaisons avec marqueurs et un environnement intérieur contrôlé. Aujourd'hui, les dernières méthodes d'estimation de posture s'appuient sur les progrès de l'intelligence artificielle basés sur l'apprentissage

machine et ne requièrent que quelques images, voir une seule de la scène et du sujet. De plus, ces méthodes fonctionnent sans marqueurs et sont utilisables avec n'importe quelle donnée visuelle provenant de caméras ou appareils photo grand public.

Cependant, ces succès restent encore relatifs pour plusieurs raisons. La précision des méthodes estimant la posture en trois dimensions reste encore bien inférieure à celle des systèmes optiques traditionnels avec marqueurs. Cette précision reste encore trop faible pour certaines analyses scientifiques. De plus, même si la contrainte des marqueurs a été levée, produisant ainsi des méthodes beaucoup moins intrusives, ces systèmes requièrent souvent un processus de calibration qui peut être long et complexe. Enfin, même si l'apprentissage profond permet une meilleure caractérisation des articulations dans les images, de nombreux biais existent inévitablement dans les jeux de données de grande échelle utilisés pour l'entraînement.

L'enjeu du choix des méthodes d'estimation de posture définira donc la capture de mouvement de demain. Les méthodes basées sur notre connaissance de la géométrie projective sont très exactes (s'il n'y a pas de bruit d'acquisition) et permettent d'obtenir les coordonnées dans une scène 3D, mais ne tiennent pas compte des contraintes apriori sur la structure du corps humain. Les modèles de contraintes permettent de respecter cette structure, mais au prix d'une spécificité aux sujets et aux scènes souvent moins importante. Enfin, les méthodes basées sur l'apprentissage produisent des résultats de plus en plus performants tant en précision qu'en robustesse, mais nécessitent toujours plus de données pour généraliser leurs performances.

1.3 Questions de recherche

L'ambition à long terme est la conception d'un outil robuste et peu coûteux pour la capture en continu des mouvements humains. Ce travail de thèse doit permettre de mettre en place un cadre conceptuel, logiciel et matériel qui démontre la faisabilité d'un système de monitoring ubiquitaire des mouvements humains (à domicile, au travail, etc.) permettant d'interroger concrètement le souhaitable et les usages de ce type de monitoring. Les applications d'un tel outil sont larges : prévention-santé (par exemple pour la détection des troubles musculo-squelettiques), l'analyse des comportements individuels ou sociaux (par exemple les postures, la communication non verbale), l'interface domotique...

La question de recherche qui structure cette thèse est la suivante : **Quel cadre conceptuel, quelle architecture logicielle, et quels capteurs/systèmes peut-on combiner pour mettre à disposition de la communauté scientifique un système de capture des mouvements humains qui soit à la fois : sans marqueurs, robuste, exact, à bas coût et dans une zone de la taille d'une pièce ou d'une habitation ?**

Cette problématique introduit un ensemble de spécifications techniques pour lesquelles il existe déjà des éléments de réponse partiels dans la littérature (exactitude) et d'autres qui sont moins explorées (robustesse et ergonomie des méthodes). Un état de l'art complet des méthodes de capture de mouvement et des algorithmes d'estimation de posture 3D a guidé le choix des capteurs, du système d'acquisition et des algorithmes utilisés. Plus précisément, il répond à la question : **Quelle exactitude peut-on attendre des différents systèmes et méthodes d'estimation de posture sans marqueurs ? Et quelles contraintes et quels prérequis introduisent-elles ?**

La réalisation d'un prototype fonctionnel a permis de confronter les choix issus de cet état de l'art pour vérifier si la performance théorique du système est généralisable. Ce prototype a pour but de démontrer : **Est-il possible de réaliser un système complet et fonctionnel de capture du mouvement sans marqueurs à bas coût ?** L'exactitude du système est étudiée en fonction de la méthode d'estimation de posture utilisée à des fins d'amélioration et de comparaison avec l'état de l'art. La robustesse et l'ergonomie sont discutées pour proposer des pistes permettant le développement et le déploiement du système.

Enfin, la nature tridimensionnelle du problème et des données produites par le prototype ont également entraîné un questionnement sur la meilleure façon de les exploiter. La question de recherche à laquelle nous avons tenté de répondre est la suivante : **Est-il possible de fusionner des caractéristiques visuelles avec une connaissance apriori de la géométrie d'une scène en 3D afin d'obtenir une plus grande exactitude lors de l'estimation de posture ?**

1.4 Plan de la thèse

La structure de ce manuscrit est la suivante :

- L'état de l'art de la thèse :
  - Le chapitre 2 présente le contexte scientifique et technologique des travaux de cette thèse. D'abord, les motivations pour la conception et la mise en place d'un prototype de système de capture du mouvement humain sont explicitées. Sont ensuite décrits les différents critères et spécifications nécessaires à l'évaluation d'un tel système comme l'*exactitude* et la *robustesse*. Le coût du matériel nécessaire est également abordé, et mis en parallèle avec les deux autres critères de performance précédents. Une définition des tâches d'estimation de posture 2D et 3D en vision par ordinateur est rappelée. Enfin, les concepts d'apprentissage machine, de géométrie projective et reconstruction 3D mobilisés pour le prototype final sont répertoriés et définis dans cette partie. Les différentes autres options qui existent pour la capture du mouvement optique et non optique sont également présentées.
  - Le chapitre 3 décrit l'ensemble des méthodes d'estimation de posture 3D de l'état de l'art et les compare sur la base de trois critères de performance (*exactitude*, *robustesse* et *vitesse*). Il permet d'orienter le choix du ou des algorithmes utilisés pour le prototype.
- Les différentes contributions et le prototype :
  - Le chapitre 4 aborde les essais effectués sur la fusion des caractéristiques 2D et 3D obtenues à partir de prétraitements ou au bout d'une première étape de détection 2D dans le cadre d'un système multivues.
  - Le chapitre 5 décrit le prototype que nous avons développé et l'évalue et présente les choix matériels et algorithmiques pour la mise en place de ce type de système à l'avenir. Une comparaison avec les méthodes d'estimation de posture et les autres systèmes de capture du mouvement complets est faite sur la base des critères de performances spécifiés.
- Le chapitre 6 fait le bilan pour la mise en œuvre du prototype en tant que système complet et fonctionnel. Les perspectives scientifiques au niveau de l'acquisition, mais aussi de la reconstruction et de l'estimation de posture sont abordées. Enfin, les limitations qu'amène un tel système seront discutées.

Première partie

Etat de l'art

Chapitre 2

Contexte Scientifique

Contents

2.1 Motivations

La capture du mouvement humain est un défi technologique qui anime les chercheurs depuis les balbutiements de la photographie. Ces applications sont nombreuses tout comme les technologies utilisées pour la réaliser. Le chapitre suivant s'attache à la définir au mieux ainsi qu'à décrire l'état des connaissances nécessaires à son fonctionnement dans divers domaines scientifiques.

Pour les contributions de cette thèse, les applications ciblées concernent l'étude du mouvement humain. Un exemple d'utilisation peut être l'étude de la marche ou des mouvements des membres antérieurs. Ce type d'étude est souvent effectué dans un cadre médical sur des patients à des fins diagnostiques ou d'analyse. Dans certains cas, les systèmes de capture de mouvement commerciaux actuels ne sont pas adaptés et c'est ce qui conduit à la conception d'un nouveau système plus flexible, moins intrusif.

Ces contraintes entraînent des choix de conception qui constituent les spécifications du prototype présenté dans ces travaux. Plusieurs axes ou critères de performances peuvent influencer ces choix de types de capteurs et algorithmes utilisés pour la capture de mouvement. Ainsi, la section qui suit dans ce chapitre commence par le détail de ces critères et les applique aux spécifications du prototype.

Une fois le contexte théorique et les spécifications définis, une synthèse comparative des méthodes actuelles d'estimation de posture est proposée. En découle le choix de l'estimation de posture à partir d'images sous différents angles. Les contributions présentées dans la partie II sont guidées par cette analyse. Celles-ci sont doubles, des travaux de recherche sur la fusion de donnée pour déterminer la méthode exploitant au mieux les images multivue 4 et la réalisation du prototype et son évaluation 5.

2.2 Spécifications du système et critères d'évaluation

La question de recherche principale de la thèse (1.3) introduit plusieurs spécifications ou caractéristiques pour le système qui doit être développé. Ces caractéristiques sont les suivantes : **Sans-marqueurs, robuste, exacte, à bas coût, taille d'une pièce**

Ces différents critères proviennent de besoins formulés par des médecins et chercheurs en science du mouvement humain. De nombreuses études sont en effet réalisées avec des systèmes traditionnels de motion capture qui, bien que très précis, ne sont pas utilisables dans certains contextes (trop intrusifs, trop coûteux et peu flexible en terme d'installation). Dans ces cas, l'utilisation du capteur Kinect de Microsoft a été testée, mais possède aussi ses propres contraintes (précision moindre, confusion entre plusieurs personnes). Ce sont ces retours d'expérience qui ont principalement motivé ce projet. Il convient cependant de bien définir ces attentes pour le système, ce qui fera l'objet de cette section.

Moeslund and Granum (2001) définissent 3 critères de performance pour les systèmes d'estimation de posture humaine : L'exactitude, la robustesse et la vitesse. La robustesse est définie comme la capacité d'un système à fonctionner sous un nombre plus ou moins important de contraintes (par rapport aux types de mouvement, à l'apparence du sujet ou de la scène). Il est donc possible de rassembler certaines spécifications attendues pour le système voulu dans la catégorie de la robustesse. Nous organiserons donc ces critères en 3 catégories avec, en plus de la robustesse, l'exactitude et le coût total du système. La vitesse étant moins importante dans le cadre d'analyses scientifiques ou médicales qui peuvent être réalisées hors ligne, elle ne sera pas considérée pour évaluer le prototype final.

2.2.1 Critères de robustesse

L'analyse du mouvement humain utilise régulièrement la *motion capture* avec marqueurs avec succès pour obtenir la position des articulations avec précision. Cependant, ces dispositifs sont très contraignants pour le sujet qui doit revêtir une combinaison et des marqueurs réfléchissants sur des points clefs pour localiser ses membres et articulations. Cela pose un problème notamment dans l'analyse du mouvement pathologique ou pour des personnes en perte de mobilité. La contrainte la plus importante à dépasser et donc l'utilisation de ces marqueurs pour obtenir une méthode **sans marqueur** la plus fiable possible.

Dans la définition initiale du sujet, la "robustesse" fait référence à deux contraintes que décrivent déjà Moeslund and Granum (2001). Ces contraintes sont les suivantes : une scène éclairée de manière uniforme et la présence de plusieurs personnes dans

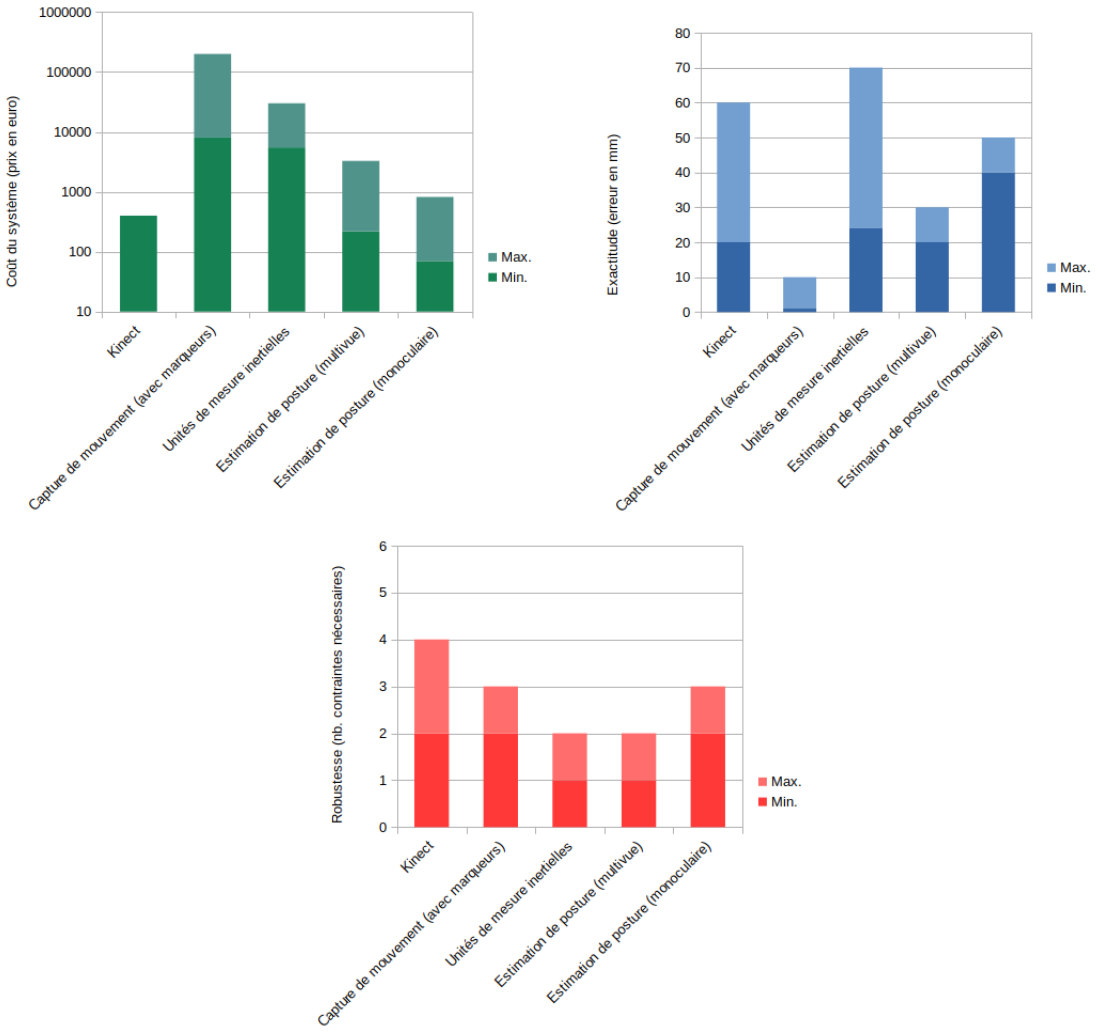


FIGURE 2.1 : Comparatif des systèmes permettant la capture du mouvement humain. **Prix** entre systèmes "minimaux" et "maximaux" : pour 6 où 10 caméras infrarouges pour capture la capture de mouvement : 3DMA STT et Qualisys Arqus ; estimation de posture : module caméra raspberry et caméras GigE (x4 pour le multivue) ; IMU : x10 unités de Output capture et Xsens Awinda. **Exactitude** en erreur moyenne sur l'ensemble des articulations suivies. **Robustesse** basée sur six contraintes principales : Marqueurs, éclairage en lumière uniforme, multipersonne, volume de capture réduit, position relative et étape de calibration nécessaire.

le volume de capture. La capture de mouvement optique est historiquement difficilement réalisable à la lumière du soleil ou si la pièce contient des objets fortement réfléchissants autres que les marqueurs. Il existe depuis la possibilité de filtrer les signaux dans ce type d'environnement directement sur l'optique des caméras ou en post-traitement (Hatfield and Sheirman, 2010). Le futur système doit, lui, pouvoir être utilisé en situation réelle sous **tout type d'éclairage**, y compris en lumière naturelle. Dans une moindre mesure, le système doit pouvoir capturer le mouvement de plusieurs personnes sans confusion pour permettre l'analyse des mouvements multipersonne.

Enfin, la nature de la zone de capture est le dernier critère important. La plupart des études menées en laboratoire ou dans un cadre médical ne nécessitent pas des zones de capture de plus de 20 m^2. Une exception serait pour l'étude de la marche qui demande de plus grandes surfaces. Cette nécessité du système de pouvoir **capturer un volume de taille moyenne** limite en partie l'utilisation de méthode qui se basent sur un seul capteur. Enfin, ces méthodes dites "monoculaires" basée sur des images ou vidéos ne permettent pas d'obtenir une **position absolue** du sujet dans la scène qui est souvent importante pour analyser le sujet en interaction avec son environnement. Pour l'obtenir, il est nécessaire de déterminer la position dans la scène d'au moins une articulation (dite "racine") pour y localiser la posture complète. Dans le cas du multivues vidéo, mais aussi de la capture de mouvement avec marqueurs, il est nécessaire de procéder à la **calibration** de l'ensemble du système de caméras pour reconstruire les points d'articulation en trois dimensions.

Ces cinq critères permettent d'évaluer la robustesse générale spécifique au type d'application scientifique ou médicale. La figure 2.1 représente la robustesse en fonction du nombre de ces critères que les différents systèmes peuvent respecter (un critère est considéré comme à moitié respecté s'il nécessite des conditions bien précises ou si seulement une partie des systèmes du type en question le respectent). La capture du mouvement classique respecte les contraintes du volume de capture et de la scène, mais nécessite des marqueurs et ne fonctionne pas avec de la lumière naturelle. Elle peut permettre la reconnaissance de plusieurs personnes en simultané, mais est sujette à la confusion des marqueurs dans la scène. C'est l'inverse pour les méthodes monoculaires qui fonctionnent sans marqueurs, en lumière naturelle, mais permettent difficilement une capture d'un volume de capture moyen ou large. Kinect a la même contrainte par rapport à la lumière naturelle (à cause du capteur de profondeur) mais permet la détection multipersonne malgré des confusions fréquentes. Pour finir, les méthodes multivues fonctionnent avec presque toutes les contraintes à l'exception du multipersonne (variable selon la méthode) et de la nécessité d'une étape de calibration. Il est également nécessaire

2.2.2 Critère d'exactitude

Les systèmes de capture du mouvement basés sur des marqueurs suivent les marqueurs lumineux avec une erreur moyenne de 1 à 2 mm. Cependant, l'erreur par rapport à la position exacte des articulations peut excéder 10 mm à cause du mouvement des marqueurs à la surface du corps d'après Colyer et al. (2018). S'ajoute à cela l'erreur de mesure due au placement et à la confusion des marqueurs entre eux lors de l'acquisition. Cette erreur tend néanmoins à s'amoindrir sur les systèmes récents. Il est donc d'usage de considérer les mesures issues de la capture du mouvement avec marqueur comme référence en termes de précision. En effet, les méthodes "gold-standard" permettant une mesure très proches de la position réelle des articulations, comme les broches intracorticales ou la fluoroscopie, sont rarement utilisées du fait de leur nature fortement intrusive (Colyer et al., 2018). L'exactitude pour les méthodes basées sur des images est donc souvent calculée à partir d'une vérité terrain issue de la capture de mouvement avec marqueurs.

Pour les méthodes basées sur des images et vidéos, l'erreur 3D moyenne (voire 3.2.1) est plus importante, et varie entre 50 mm pour les méthodes monoculaires (Wandt and Rosenhahn (2019), Kocabas et al. (2019b)) avec les meilleurs atteignant 41 mm (Kolotouros et al., 2019). Les méthodes basées sur des vidéos tendent à obtenir une exactitude autour de 45 à 40 mm (Pavllo et al. (2019), Chen et al. (2020) et Wang et al. (2020)).

Le capteur Kinect a également été utilisé pour l'estimation de posture humaine (Shotton et al.). Cependant, il est difficile de déterminer son exactitude en l'absence de jeu de donnée à grande échelle contenant des données acquises avec ces capteurs. Faity et al. (2022) proposent une évaluation du Kinect v2 pour le mouvement de la partie supérieure du corps. Le déplacement des articulations du tronc, de l'épaule, du coude et de la main y est comparé avec un système de capture du mouvement classique (Vicon). L'erreur quadratique moyenne reportée pour l'ensemble de ces articulations atteint 38,95 mm (avec 20 mm pour le tronc, mais 60 mm pour le coude). Ce résultat avoisine les meilleures méthodes d'estimation de posture monoculaire évaluée sur les jeux de données référence en vision par ordinateur. Enfin, les méthodes multivue permettent d'obtenir des scores d'erreur variant entre 30 à 20 mm (Iskakov et al. (2019), He et al. (2020)). Le chapitre 3 décris plus en détails ces méthodes.

2.2.3 Coût du système

Les prix des différents systèmes représentés (fig. 2.1) et susceptible d'évoluer. Cependant, il reste encore une différence significative due au matériel utilisé. La motion capture avec marqueur, bien que tendant à se démocratiser, reste la solution la plus

coûteuse. En effet, un unique capteur dans ces systèmes peu coûter plusieurs milliers d'euros. Un système complet avec 10 caméras peut atteindre de 100 000 à 200 000 euros pour les versions les plus récentes des systèmes dans le commerce. Pour un système avec un volume de capture moyen (4x4 m pour 3 m de hauteur) 6 caméras peuvent suffire et on atteint autour de 40 000 euros.

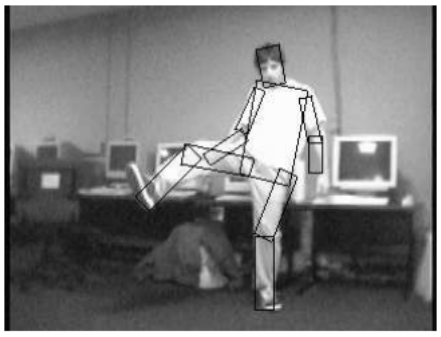
Le système Kinect v2 n'est plus vendu ni maintenu par Microsoft, mais son coût avoisinait la centaine d'euros. Les systèmes les moins coûteux se basent sur des images ou vidéos provenant d'un unique capteur. Les résultats obtenus sur la plupart des méthodes l'état de l'art fonctionnent avec des images de résolution moyenne (comme l'attestent les images des jeux de données de référence comme Ionescu et al. (2014) de 1000x1000 pixels). Ainsi, ces méthodes ne nécessitent pas des caméras très haute définition et sont donc les moins coûteuses.

L'ensemble de ces critères sont rassemblés dans la synthèse présentée dans la section 2.6 qui aborde les différents capteurs utilisables pour la capture du mouvement.

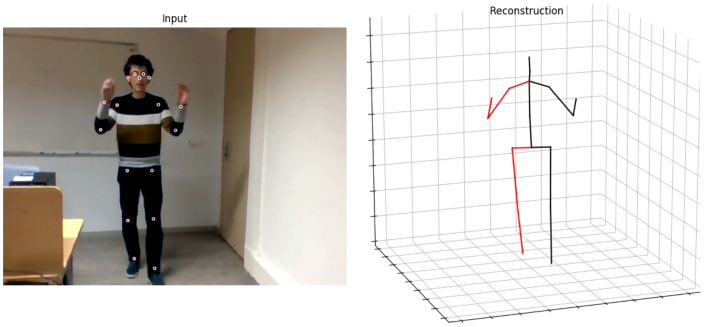
2.3 La tâche d'estimation de posture humaine

2.3.1 Définition et taxonomie

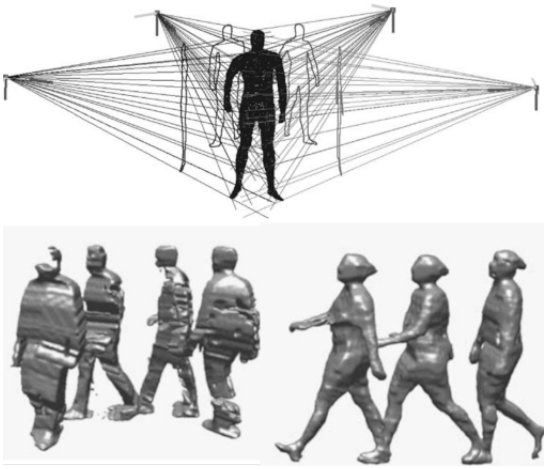
Il est possible de définir la tâche d'estimation de posture humaine en vision par ordinateur comme suit : la localisation de la posture de sujets humains et sa représentation numérique à partir d'images. Cette localisation peut être effectuée soit en deux dimensions directement dans les images, soit en trois dimensions relativement à la position du ou des capteurs par rapport à la scène capturée. Les représentations utilisées sont variées et peuvent être des formes géométriques en deux ou trois dimensions représentant les différents membres ou articulations d'une personne (comme les modèles *"mixture of parts"* voir figure 2.2). Parfois la surface entière du corps sous forme de maillages est utilisée comme représentation, on parle alors parfois de d'estimation de la forme ou d'estimation de posture dense.



(a) Rep. par boîtes
(Felzenszwalb and Huttenlocher, 2005)



(b) Rep. par squelette
(obtenu avec la méthode de Pavllo et al. (2019))



(c) "Visual hulls"
(Corazza et al. (2006) et Corazza et al. (2009))

(d) Rep. volumétrique SMPL
(Loper et al., 2015)

FIGURE 2.2 : Différentes représentations de la posture humaine.

Parfois des représentations intermédiaires sont utilisée pour estimer une posture finale : c'est souvent le cas des "*visual hulls*" qui permette d'obtenir une approximation du volume du corps d'une personne dans une installation multivue à partir de la technique de "*shape-from-silhouette*" (Laurentini, 1994). La représentation la plus couramment utilisée en vision par ordinateur est cependant la représentation "squelettique" (*"skeletton-based"*) qui décrit des points d'articulations clefs reliés par des segments. C'est une représentation simple qui permet une comparaison facile avec les méthodes de capture du mouvement avec marqueurs.

La grande diversité des méthodes proposée, mais aussi des applications qui utilisent l'estimation de posture, rends difficile la tenue d'une taxonomie exhaustive. Le plus souvent, trois critères sont utilisés pour classer ces méthodes (voir figure 2.3) : Les représentations (comme décrit plus haut), le type de donnée nécessaire en entrée

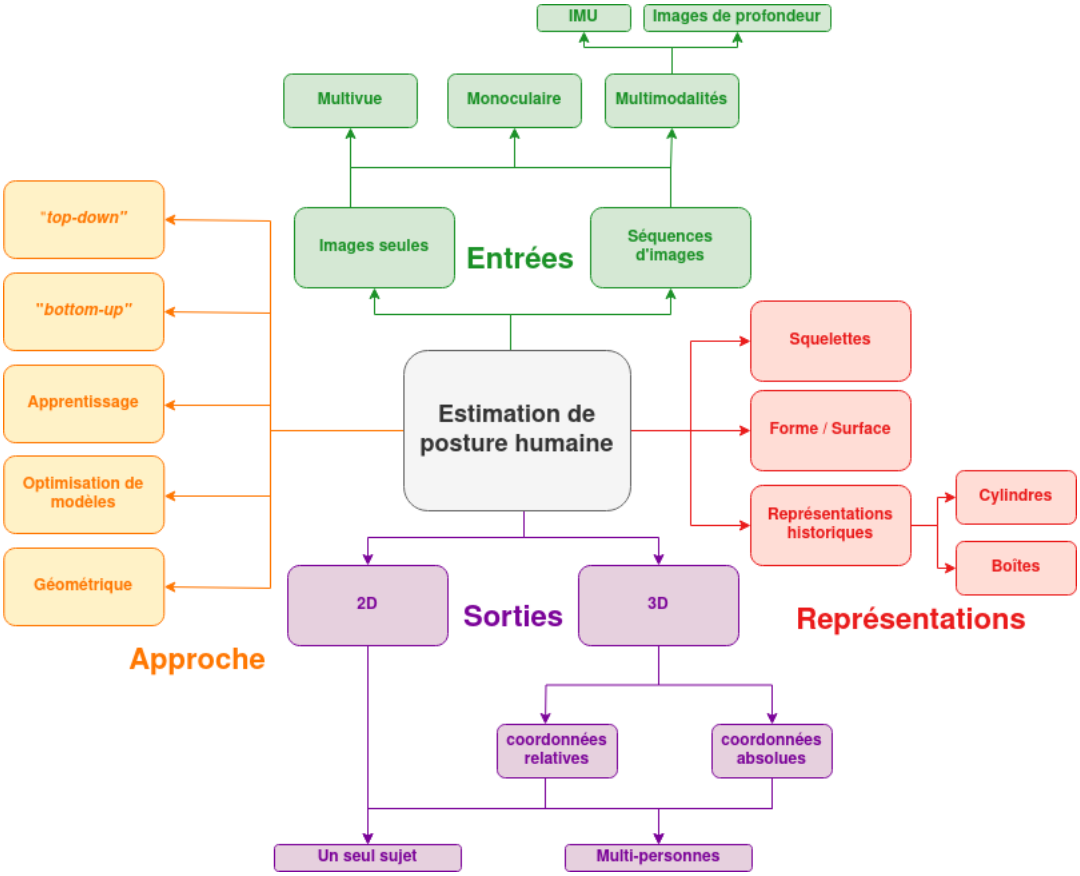


FIGURE 2.3 : Taxonomie de l'estimation posture humaine en vision par ordinateur.

et enfin la nature des informations produites en sortie. À un niveau d'abstraction plus haut, vient s'ajouter le type d'approche utilisée. Plusieurs méthodes émanant des divers champs de recherche sont utilisés et parfois combinés pour l'estimation de posture (optimisation, apprentissage machine, statistique, géométrie, etc.).

Il est possible de trouver dans la littérature les termes *"bottom-up"* (montante) et *"top-down"* (descendante) pour qualifier certaines méthodes. Dans le contexte où sont détectés plusieurs sujets dans les images (multipersonnes), l'approche *"top-down"* nécessite dans un premier temps un détecteur de personne qui produit les boîtes englobantes pour chaque sujet dans l'image. La posture humaine est ensuite déterminée pour chaque personne individuellement. À l'inverse, l'approche *"bottom-up"* extrait les éléments de la posture (point d'articulations) dans toute l'image avant de classifier leur appartenance à chaque sujet. Pour l'estimation de posture 3D, et plus généralement, *"top-down"* qualifie les méthodes qui se basent sur des représentations intermédiaires simples (posture 2D) avant l'estimation finale en trois dimensions. Elles s'opposent aux méthodes *"bottom-up"* qui utilisent des modèles plus complexes du corps humain

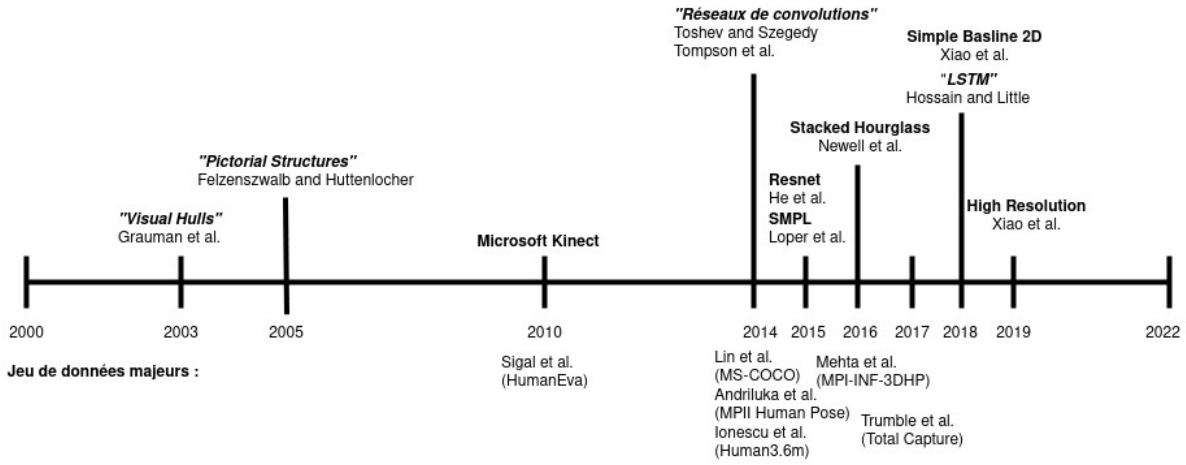


FIGURE 2.4 : Chronologie des méthodes d'estimation de posture humaine. Les articles introduisant des jeux de données majeurs ou de nouvelles approches sont reportés.

auxquels on fait correspondre aux mieux les données des images (parfois aussi appelée *"model-based"* contre *"model-free"*). Certains systèmes intègrent plusieurs approches successivement ou en parallèle (ex : Kolotouros et al. (2019) utilisent un réseau de neurone et le modèle SMPL conjointement).

Ces quatre critères de comparaison sont cependant fortement dépendants entre eux : par exemple, certaines représentations ne sont pas adaptées à l'estimation de posture 2D ou encore certaines approches nécessitent des données précises (méthodes temporelles et données vidéos). La complexité et la diversité de la tâche d'estimation de posture nécessite donc une grande attention lorsque les résultats des différentes méthodes sont comparés.

2.3.2 Historique

Historiquement, les premières approches pour estimer la posture humaine utilisaient des caractéristiques prédéfinies extraites des images. Ces caractéristiques étaient utilisées pour adapter des modèles de contraintes du corps humain ou pour entraîner des algorithmes d'apprentissage (par exemple, Mori et al. (2004) extraient des caractéristiques de formes, luminance et focus pour chaque membre du corps et les combinent pour entraîner un modèle de régression logistique).

En 2004, Grauman et al. (2003) introduisent utilisent les *visual hulls* pour obtenir l'une des premières représentations volumétriques de la posture. En 2005, Felzenszwalb and Huttenlocher (2005) appliquent les *"pictorial structures"* ou *"part based mo-*

del" de Fischler and Elschlager (1973) à la posture humaine. Ces modèles considèrent différentes parties (membres ou articulations) qui peuvent être reliées par paires. Le principe de la méthode est de minimiser une fonction d'énergie qui dépend de la correspondance aux données de l'image des parties ainsi que de la cohérence avec un modèle déformable du corps.

Suite à la réussite sur les challenges d'Imagenet (Deng et al., 2009) en 2012 d'AlexNet (Krizhevsky et al., 2017), l'utilisation d'architectures de réseaux de neurones convolutifs se popularise pour de nombreuses tâches en vision par ordinateur. Pour l'estimation de posture humaine, c'est en 2014 que Toshev and Szegedy (2014) et Tompson et al. (2014) proposent des architectures utilisant des réseaux de neurones. Cette période coïncide avec la publication de jeux de donnée à grande échelle contenant des images de personnes, qu'ils soient généralistes (MS-COCO : Lin et al. (2015)) ou spécialisés (Human3.6M : Ionescu et al. (2014)). Depuis, les architectures de pointes du domaine sont adaptées et combinées avec les approches classiques pour l'estimation de posture 2D et 3D.

2.4 Apport de l'apprentissage profond

Cette section énumère les différents types d'architectures de réseau de neurones artificiels utilisés pour l'estimation de posture et en décris le fonctionnement général. De plus, pour chacune d'elle, un exemple d'utilisation pour l'estimation de posture est présenté. Suit un descriptif plus général des différentes approches que les chercheurs du domaine choisissent de prendre lors de la conception de ces réseaux pour l'estimation de posture. Enfin, les limites que pose ce type de méthodes seront brièvement abordées.

2.4.1 Description des architectures

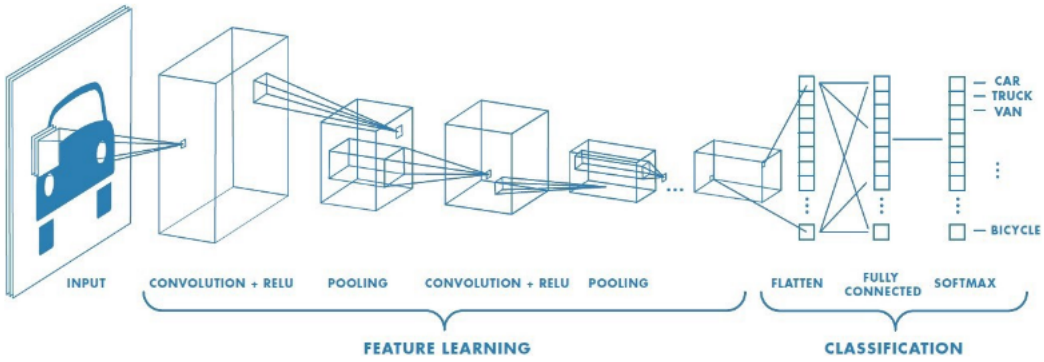
Les architectures de **réseaux de neurones convolutifs** sont composées de couches dont les plus répandues sont : convolutions, *pooling* et dense (ou totalement connectées) (voir figure 2.5 (a)). Elles nécessitent également une fonction d'activation après chaque couche de convolution ou dense. Dans une couche convolutive, on applique successivement des filtres de convolution à toutes les zones de l'image, suivie d'une fonction d'activation (pour permettre la rétropropagation) afin de produire plusieurs cartes d'activation (ou cartes de caractéristiques). Lors de l'apprentissage, les poids et biais appliqué lors des convolutions sont ajustés. Classiquement, ces couches sont suivies de couches qui réduisent la dimension de ces cartes d'activations (dîtes de *"pooling"*). Elles permettent l'exploitation de caractéristiques moins sensibles aux translations dans l'image. Enfin, en dernière étape, les cartes d'activations finales sont transformées en vecteurs et passé dans une série de couches "denses". Ces couches sont similaires à

un perceptron multicouches (décris plus bas) : une matrice de poids et un biais, puis une fonction d'activation sont appliquées au vecteur de caractéristiques, de manière similaires aux couches de convolutions. Les neurones (poids + activation) de chaque couche dense sont connectés à chaque neurone de la couche suivante. L'exemple le plus classique de ce type d'architectures sont les réseaux LeNet (Lecun et al., 1998) et AlexNet (Krizhevsky et al., 2017), respectivement utilisés pour des tâches de classification pour des chiffres manuscrits et pour le challenge ImageNet de 2012 (Deng et al., 2009) (voir figure 2.5 (c)). Pour l'estimation de posture, Toshev and Szegedy (2014) utilisent une architecture de réseau presque identique à celle d'AlexNet appliqué à la régression des coordonnées d'articulation directement dans l'image.

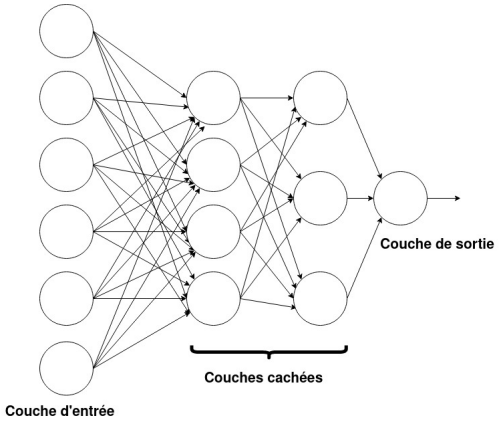
Un **perceptron multicouche** (voir figure 2.5 (b)) est un réseau de neurone artificiel où chaque neurone est relié à l'ensemble des sorties de la couche précédente (ou aux données en entrée). Le réseau applique à ces données des biais et des poids puis on les agrège avant de les passer à une fonction d'activation. Lors de l'apprentissage, une fonction de coût permet de pénaliser les neurones de la couche de sortie en fonction du résultat qui aurait été correct (à partir de l'étiquette de la sortie voulue). La procédure de rétropropagation permet de réajuster les poids de tous les neurones précédents successivement en fonction de chaque nouvelle donnée présentée au réseau. Les perceptrons multicouches sont moins utilisés lorsque les données étudiées sont des images pour lesquelles les réseaux de convolution produisent des résultats souvent meilleurs. Les couches de sorties de certains réseaux dites "totalement connectées" se rapprochent de leur fonctionnement. Cependant, Martinez et al. (2017) proposent une architecture simple très similaire au fonctionnement d'un perceptron multicouche. Les données utilisées en entrée ne sont plus des images, mais des coordonnées 2D extraites des images par des estimateurs 2D de l'état de l'art. Ainsi, ces coordonnées peuvent être vectorisées et utilisées comme base d'apprentissage pour le réseau.

Les **réseaux de neurones résiduels** ou "Resnet" (voir figure 2.6) proposés par He et al. (2015) se basent sur le principe des sauts de connexion (*"skip-connection"*). L'idée est de sauter plusieurs couches et de relier une partie du réseau avec des couches plus profondes. Cette technique permet de réduire le problème de disparition du gradient pendant l'apprentissage tout en l'accélérant, ce qui a ouvert la porte à des réseaux de plus en plus profonds (jusqu'à une centaine de couches). Les variantes de ce réseau sont appelées Resnet50, Resnet101 ou encore Resnet152 en fonction de leur nombre de couches. Ces architectures ont été utilisées avec succès, notamment en gagnant le challenge Imagenet (Deng et al., 2009) en 2015.

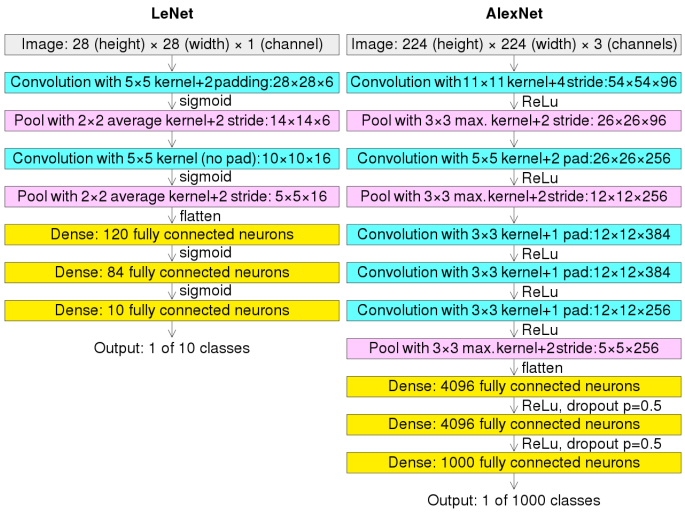
Les algorithmes d'estimation de posture ont fait une utilisation importante de ce type d'architecture avec sauts de connections. C'est principalement le cas pour l'estimation de posture 2D pour laquelle de nombreuses architectures ont été proposées



(a) Exemple d'un réseau de neurone convolutif (Aussel et al., 2019)



(b) Schéma d'un perceptron multicouche



(c) Comparaison entre LeNet (Lecun et al., 1998) et AlexNet (Krizhevsky et al., 2017) Wikimedia Commons (2021)

FIGURE 2.5 : Réseaux de neurones convolutifs et perceptron multicouches

avec comme bloc constitutif principal des "blocs résiduels". Les plus répandues sont : *"Stacked Hourglass Networks"* (Newell et al., 2016), *Cascaded Pyramid Networks* (Chen et al., 2018) et *"High Resolution Networks"* (Sun et al., 2019). De plus, les variantes de Resnet sont parfois directement entraînées pour l'estimation de posture en tant qu'architecture principale. Ces architectures (décrites plus en détails dans le chapitre suivant en 3.3.3) ont régulièrement participé à l'augmentation de précision des estimateurs de posture 2D sur des jeux de données de références pour la tâche.

Les **réseaux de neurones récurrents** (RNN), **LSTM** (*Long Short-Term Memory*) ou encore les **réseaux convolutifs temporels** (TCN) ont été conçus pour traiter des données séquentielles comme les séries temporelles, le langage naturel. Ce type d'architecture a été adapté pour l'estimation de la posture dans les vidéos. Les réseaux neuronaux récurrents ou RNN (voir figure 2.7 (a)) constituent un moyen de traiter les séquences de postures. Plus précisément, les architectures de type "*Long Short-Term Memory*" (LSTM) ont été utilisées avec succès sur des séquences d'articulation en 2D (par exemple, Hossain and Little (2018)) et d'autres sources de données comme les unités de mesure inertielles (par exemple, Trumble et al. (2017)). Cette technique a montré son succès dans la traduction de textes et d'autres tâches pour traiter des séquences avec des dépendances à long terme. À partir d'une séquence, une LSTM peut en produire une autre tout en conservant des informations sur les entrées précédentes passées (modèle séquence-à-séquence). La principale différence entre les LSTM et les RNN est l'état de leurs cellules qui est mis à jour par différentes opérations linéaires appelées "portes". Ces opérations sélectionnent et mettent à jour les informations qui sont utiles à la mémorisation (ce site Web de Christopher Olah détaille le fonctionnement des LSTM).

Le réseau convolutif temporel (voir figure 2.7 (b) et (c)) est une autre solution proposée par Bai et al. (2018) pour traiter des données séquentielles avec des dépendances à long terme et résoudre le problème d'explosion ou disparition du gradient que les réseaux récurrents peuvent avoir. Un TCN est un réseau totalement convolutif qui utilise

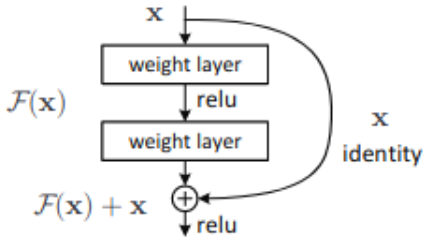
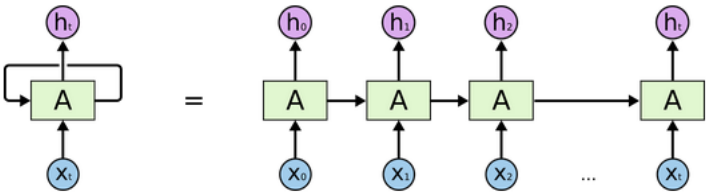
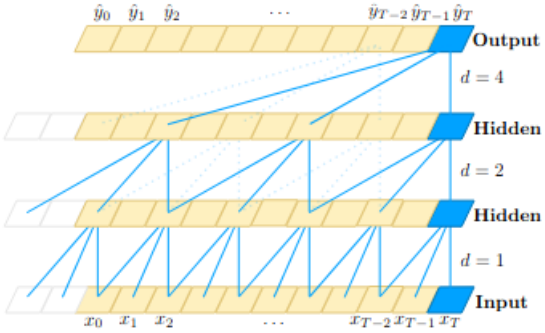


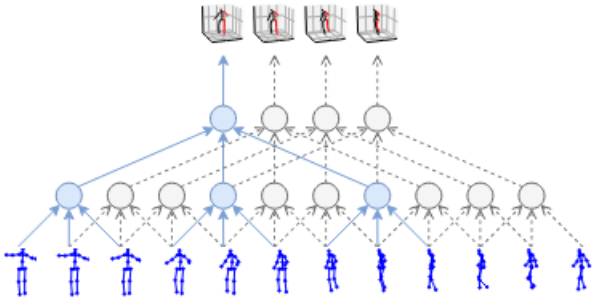
FIGURE 2.6 : Bloc résiduel illustrant le principe des sauts de connexion. Tiré de He et al. (2015)



(a) Réseau de neurone récurrent "déplié". C.Olah



(b) Exemple de convolutions causales dilatées (facteur de 1, 2 et 4). Bai et al. (2018)



(c) TCN appliqué à l'estimation de posture. Pavllo et al. (2019)

FIGURE 2.7 : Réseaux de neurones artificiels pour l'analyse de séquences

des convolutions "causales" et "dilatées". C'est-à-dire qu'elles ne prennent en compte que les informations du passé par rapport à la séquence traitée et avec un pas prédéfini permettant de considérer les informations plus anciennes de la séquence dans son champ récepteur. Pavllo et al. (2019) et Cheng et al. (2020) utilisent ce type de réseau pour analyser des vidéos contenant des séquences de postures humaines.

Le **mécanisme d'attention** vise à concentrer l'apprentissage sur les informations les plus importantes dans les données d'entrée ou les caractéristiques intermédiaires. Dans le cadre de l'estimation de posture, Cai et al. (2019) utilisent l'attention pour sélectionner les images qui contribuent le plus à l'estimation, tandis que He et al. (2020) décrivent un module *"Epipolar Transformer"* qui tire parti du multivue pour se concentrer sur l'apprentissage le long des lignes épipolaires : Les caractéristiques de l'image appariée sont fusionnées le long des lignes épipolaires correspondant aux points de jonction.

Enfin, un autre axe de recherche intéressant concerne les réseaux antagonistes génératifs (GAN) et **l'apprentissage adverse** (Goodfellow et al. (2014)). Ils ont été principalement utilisés pour l'estimation de la posture 3D de deux manières : mise en correspondance non supervisée de posture 2D à 3D (Kudo et al., 2018) et plus fréquemment comme module de validation de la posture (Cheng et al. (2020), Wandt and Rosenhahn (2019), Kocabas et al. (2019a)). Dans ce cas, un réseau discriminant amé-

liore la cohérence des postures en étant entraîné à distinguer les postures générées de celles extraites directement des vérités terrain. Une composante de "coût adverse" est ensuite propagée au réseau générateur qui estime les postures humaines 3D à partir d'informations visuelles ou de poses 2D. Ainsi, les postures qui ne sont pas cohérentes avec les configurations connues sont pénalisées.

2.4.2 Conception d'architectures pour l'estimation de posture

Les approches mises en place pour l'entraînement de réseaux de neurones artificiels pour la tâche d'estimation de posture sont principalement guidés par les jeux de données disponibles et surtout par leur étiquetage. Cet étiquetage est principalement obtenu de deux manières différentes. Tout d'abord, il est réalisé sur des images seules labellisées par des humains, produisant des jeux de données principalement utilisés pour les approches monoculaires et pour la posture 2D dans les images. L'autre possibilité est l'utilisation de la capture du mouvement avec marqueurs conjointement avec la capture vidéo d'un sujet en mouvement sous plusieurs angles (multivue). Ce type de jeu de donnée peut aussi être exploité pour la posture 3D à partir des données multivue et pour l'analyse des séquences d'images de la vidéo. Dans les deux cas, les jeux de données principaux donnent souvent des étiquettes sous forme de points d'intérêt représentant les articulations (voir section 2.3). Ainsi, la majorité des architectures se basent sur cette représentation de la posture. La nature des données disponibles influe donc sur les choix d'architectures (par exemple : RNN ou TCN pour des vidéos).

Parfois, les données brutes sont transformées avant d'être utilisées pour l'apprentissage pour des méthodes divisé en plusieurs étapes. Les caractéristiques intermédiaires produites dans ce cas sont multiples : cartes de chaleur de la localisation des articulations, représentation volumétriques (cartes de voxels ou *"visual hulls"*). Pour l'estimation de posture à partir de données multivue ce sont parfois les coordonnées 2D de chaque vue qui font office de caractéristiques intermédiaires pour des réseaux ensuite entraînés à retrouver les coordonnées 3D.

Enfin, si la méthode développée vise à estimer la posture de multiples personnes dans les images, un choix important de conception concerne l'ordre entre la détection du sujet et l'estimation de sa posture (si la méthode est descendante ou montante : voir section 2.3). Si c'est une méthode descendante, un détecteur de personne est d'abord appliqué et l'estimation de posture est réalisé sur celles-ci (comme dans Papandreou et al. (2017)). Dans ce cas, il faut alors choisir le détecteur à utiliser et la procédure pour sélectionner les boîtes englobantes qu'il produit ainsi qu'une manière de gérer les occultations inter-personne. Pour les méthodes montantes, une fois la détection des points d'articulation faite, il faut mettre en place une procédure pour les regrouper selon s'ils appartiennent aux mêmes sujets (comme pour Cao et al. (2017)). Certaines

autres méthodes (Newell et al., 2017) prédisent les deux simultanément en une seule étape.

Pour conclure, la performance de ces méthodes d'apprentissage profond est lié à la qualité des jeux de données utilisés. S'ils sont issus de données de capture du mouvement avec marqueur, les vérités terrains produites sont très précises, mais sont obtenues dans un cadre contrôlé de laboratoire en lumière non naturelle. De plus, le nombre de sujets différents filmés dépasse rarement la vingtaine d'individus. Pour les données issues d'images d'internet ou labellisé manuellement, il existe une plus grande diversité de sujet et d'action même si l'étiquettage et souvent moins exact. De plus ces données sont souvent moins riches, car il n'y souvent pas de séquence vidéo ou de multiples angles de capture pour l'estimation de posture en 3D. Tous ces facteurs font que les méthodes entraînées et évaluées sur ces jeux de données sont partiellement biaisées.

2.5 Connaissances des scènes tridimensionnelles

Pour l'estimation de posture humaine tridimensionnelle, les coordonnées des points d'articulation doivent être exprimées dans le système de référence d'une caméra ou d'un repère dans la scène. Il faut alors utiliser la géométrie épipolaire et la mise en correspondance de points dans les différentes vues pour parvenir à reconstruire la posture en trois dimensions. Cette section décrit les principes généraux de la géométrie épipolaire. Ensuite est détaillée la calibration d'un système de caméra multivue afin d'obtenir leurs paramètres extrinsèques et intrinsèques. Enfin, plusieurs méthodes couramment utilisées pour reconstruire des points en trois dimensions sont présentées.

2.5.1 Modèle sténopé

Dans un premier temps, on considère que le capteur utilisé pour observer une scène suit le modèle sténopé (ou *pinhole* en anglais). Ce modèle décrit une projection de points du monde réel en trois dimensions dans le plan image sans prendre en compte de potentielles distorsions optiques. Il est alors possible de passer du repère monde au repère caméra en appliquant une rotation puis une translation. Cette matrice de rotation ($\mathbf{R}$) et ce vecteur de translation ($\mathbf{t}$) sont appelés paramètres extrinsèques de la caméra (voir équation 2.1).

$$\begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} \end{bmatrix} \begin{bmatrix} X_{mon} \\ Y_{mon} \\ Z_{mon} \end{bmatrix} + \mathbf{t} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_{mon} \\ Y_{mon} \\ Z_{mon} \\ 1 \end{bmatrix} \tag{2.1}$$

L'étape suivante consiste à transformer les points dans le repère caméra vers le repère image. Pour cela, il faut d'abord faire projection perspective vers le plan image à partir de la distance focale de la caméra utilisée (voir équation 2.2). Les coordonnées *x* et *y* obtenues sont les coordonnées dans le plan rétinien (en unité métrique) du point projeté.

Pour passer des coordonnées métriques (*x*, *y*) aux coordonnées discrètes en pixels (*u*, *v*) on utilise une dernière transformation à l'aide de la matrice (**A**). Cette matrice est constituée des paramètres suivants : u_0, v_0 les coordonnées du centre optique dans le plan image et k_u, k_v le nombre de pixels par unité de longueur dans les deux directions (ou facteur d'agrandissement). Avec $\alpha_u = fk_u$ et $\alpha_v = fk_v$.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} fx \\ fy \\ 1 \end{bmatrix}$$
$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{A} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.2}$$

On ajoute parfois γ le facteur de "non-orthogonalité" ou *"skew factor"*. La matrice **A** permet la transformation du repère monde au repère image en coordonnés. La matrice **A** est la matrice des paramètres intrinsèques de la caméra (voir équation 2.3).

$$\mathbf{A} = \begin{bmatrix} \alpha_u & \gamma & u_0 & 0 \\ 0 & \alpha_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{2.3}$$

2.5.2 Géométrie épipolaire

On considère dorénavant un système stéréoscopique (figure 2.8), c'est-à-dire au moins deux caméras observant un même point dans une scène. L'objectif est de pouvoir, par exemple, trianguler un point visible dans les images des deux caméras. Avec un seul capteur en effet, pour un point de l'image, il existe une infinité de points 3D dans la scène pouvant en être la projection. Dans un système stéréoscopique, l'ensemble de ces points correspondent dans l'autre image à une droite appelée droite épipolaire. Le

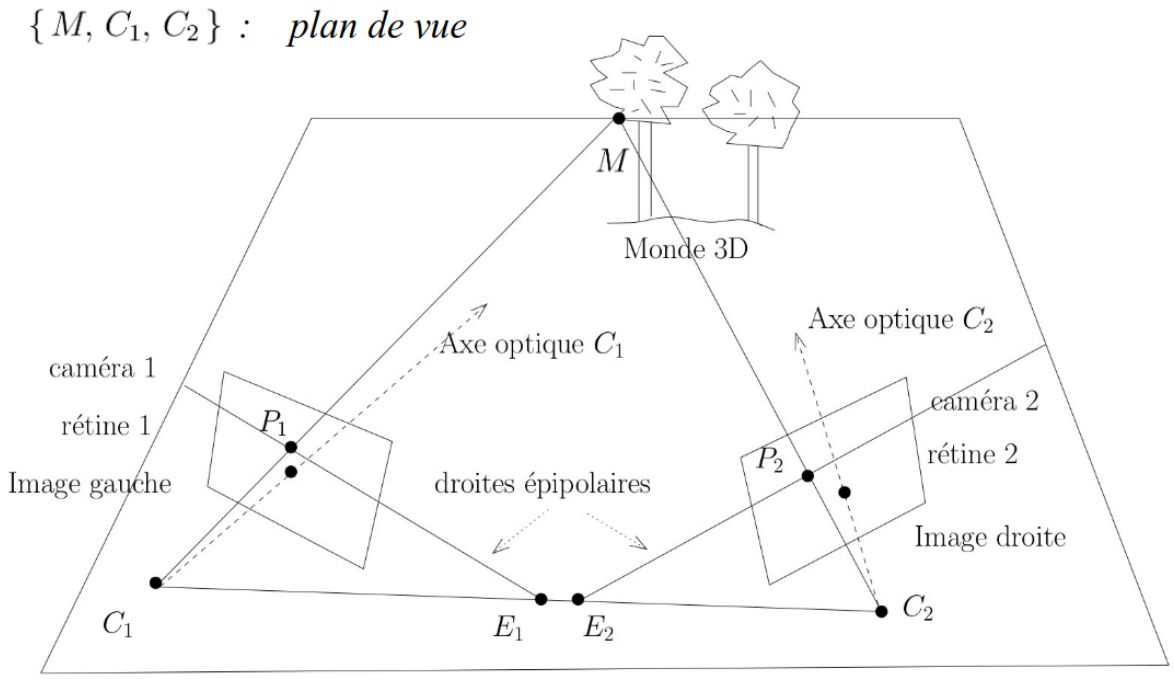


FIGURE 2.8 : Illustration de la géométrie d'un système stéréoscopique

point correspondant dans cette image au point 3D se trouve sur cette droite. Les épipôles (E1 et E2 dans la figure 2.8) sont les points d'intersection entre la droite passant par les centres de projection (C1 et C2) et le plan image.

Il est possible d'obtenir une matrice **F**, dite matrice fondamentale, qui permet d'obtenir la droite épipolaire dans une vue correspondant à un point de l'autre image pour tout point de celle-ci. Ainsi, si un point du monde M est projeté dans la première image en P1 et dans la seconde en P2, ils satisfont l'équation suivante :

$$P_1^T \mathbf{F} P_2 = 0$$

2.5.3 Calibration d'un système multivue

Avant de pouvoir trianguler ou reconstruire les points 3D, il faut d'abord obtenir les paramètres intrinsèques et extrinsèques détaillés dans les sous-sections précédentes. Les paramètres intrinsèques correspondent aux propriétés de chaque caméra et sont nécessaires pour tous les capteurs du système. Les paramètres extrinsèques correspondent aux matrices de rotations et vecteurs de translation utiles pour passer du repère d'une caméra à l'autre. Le processus de calibration décrit ici est celui proposé par Zhang (2000). Pour une description plus précise de la calibration du système mul-

tivue du prototype, voir partie II contributions, sous-section 5.3.1. La calibration suit les étapes :

- Impression d'une mire de calibration (souvent un échiquier)
- Capture de plusieurs images de la mire sous différents angles de vue
- Détection des caractéristiques dans la mire (ex : coins de l'échiquier)
- Estimation des paramètres des caméras (et du *skew-factor* le cas échéant)
- Affiner les paramètres obtenus avec l'erreur de reprojection

À partir de la 4ème étape, lorsque les points d'intérêts ont été détectés dans la mire, il est possible d'établir une homographie (**H**) qui fait la relation entre un point 2D de l'image $\tilde{\mathbf{m}}$ et un point 3D dans la scène $\tilde{\mathbf{M}}$ (voir équation 2.4). C'est une homographie de plans (matrice 3x3 en coordonnées homogènes) car la mire et la rétine sont planaires.

$$s\tilde{\mathbf{m}} = \mathbf{H}\tilde{\mathbf{M}} \text{ avec } \mathbf{H} = \mathbf{A}(Rt) \tag{2.4}$$

On pose $\mathbf{H} = \begin{bmatrix} h1 & h2 & h3 \end{bmatrix}$. De plus, si on fait l'approximation que le plan de l'objet est à Z=0 dans le repère global de la scène, il est possible d'écrire (à un facteur d'échelle λ près) :

$$\begin{bmatrix} h1 & h2 & h3 \end{bmatrix} = \lambda \mathbf{A} \begin{bmatrix} r1 & r2 & t \end{bmatrix}$$

Vu que les vecteurs *r1* et *r2* sont orthonormaux, on peut alors écrire l'équation suivante :

$$\begin{aligned} h_1^T \mathbf{A}^{-T} \mathbf{A}^{-1} h_2 &= 0 \\ h_1^T \mathbf{A}^{-T} \mathbf{A}^{-1} h_1 &= h_2^T \mathbf{A}^{-T} \mathbf{A}^{-1} h_2 \end{aligned} \tag{2.5}$$

Pour trouver une solution à ce problème, on pose **B** :

$$\mathbf{B} = \mathbf{A}^{-T}\mathbf{A}^{-1} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix}$$

B est symétrique et définie par le vecteur : $b = \begin{bmatrix} B_{11} & B_{12} & B_{22} & B_{13} & B_{23} & B_{33} \end{bmatrix}^T$. Si on pose la ième colonne de **H**, h_i on a :

$$h_i^T B h_j = v_{ij}^T b \tag{2.6}$$

Dans l'équation 2.6 le vecteur v vaut :

$$v_{ij} = \begin{bmatrix} h_{i1}h_{j1}, & h_{i1}h_{j2} + h_{i2}h_{j1}, & h_{i2}h_{j2}, & h_{i3}h_{j1} + h_{i1}h_{j3}, & h_{i3}h_{j2} + h_{i2}h_{j3}, & h_{i3}h_{j3} \end{bmatrix}$$

D'après les contraintes obtenues des homographies en (2.5) il est alors possible de réécrire (2.6) sous forme de deux équations homogènes :

$$\begin{cases} v_{12}^T b = 0 \\ (v_{11} - v_{22})^T b = 0 \end{cases} \tag{2.7}$$

Ensuite, avec n images résulte un système de n équations qui contient les équations précédentes sous forme de la matrice $\mathbf{V}$ (2n x 6) :

$$\mathbf{V}b = 0 \tag{2.8}$$

Il est alors possible de résoudre le système (2.8), au facteur d'échelle près, si l'on pose le nombre d'images utilisé n$\geq$3 et cette solution est unique. Cette solution est le vecteur propre correspondant à la plus petite valeur propre de $\mathbf{V}^T\mathbf{V}$. Par la suite, il est aussi possible de prendre en compte la distorsion radiale. Pour avoir le détail du processus, se référer à l'article original Zhang (2000). Une fois que les paramètres de caméras sont estimés, la triangulation de points images en points 3D est possible.

2.5.4 Reconstruction en trois dimensions

Il existe plusieurs algorithmes permettant la triangulation multivues. Certains se basent sur des contraintes géométriques comme l'algorithme *"midpoint"*. D'autres méthodes algébriques utilisent la transformation linéaire directe (*"Direct Linear Transform"* ou DLT).

Le principe de l'algorithme *midpoint* (Beardsley et al., 1994) est de trouver le segment le plus court entre les deux lignes de vue passant par des points en correspondances (voir figure 2.9). le point central de ce segment (point 3D) est ainsi triangulé. Pour cela, l'algorithme cherche à obtenir le point qui minimise sa distance au deux droites. Il est possible de généraliser cette méthode au multivue (par exemple Ramalingam et al. (2006)). Bien qu'aisé à calculer, le défaut principal de *midpoint* d'après Hartley and Sturm (1997) est qu'il n'est pas invariant aux transformations affines et projectives.

Dans Hartley and Sturm (1997), les auteurs proposent deux méthodes "linéaires" dont le principe est de résoudre un système d'équations linéaires provenant de la relation suivante :

$$\mathbf{x} = \mathbf{PX}$$

Avec x = w$(u, v, 1)^T$ un point image de coordonnée *u*, *v* au facteur d'échelle w près. X est le point 3D et P la matrice de calibration (3x4). Le facteur d'échelle est inconnu, mais on peut écrire :

$$\begin{cases} wu = \mathbf{P}_1^T \mathbf{X} \\ wv = \mathbf{P}_2^T \mathbf{X} \\ w = \mathbf{P}_3^T \mathbf{X} \end{cases}$$

Et, en éliminant w, l'équation obtenue devient :

$$\begin{cases} (u\mathbf{P}_3^T - \mathbf{P}_1^T)\mathbf{X} = 0 \\ (v\mathbf{P}_3^T - \mathbf{P}_2^T)\mathbf{X} = 0 \end{cases} \tag{2.9}$$

Les équations 2.9 peuvent être écrites sous la forme Ax = 0. Dans le cas de l'article d'origine, c'est un système de quatre équations linéaires et A est une matrice 4x4. Le même principe peut être appliqué en multivue avec n points image et une matrice A de taille (2n)x4.

La première manière de résoudre le système proposée dans Hartley and Sturm (1997) (désignée par les auteurs comme *"Linear-Eigen"* et parfois appelé méthode homogène). Cette solution est le vecteur propre correspondant à la plus petite valeur propre de la matrice $A^T A$ (typiquement obtenue avec la méthode de Jacobi ou la décomposition en valeurs singulières).

La deuxième manière de procéder dans Hartley and Sturm (1997) (désignée par les auteurs comme *"Linear-LS"* pour least-square, parfois appelée méthode non homogène dans la littérature) considère le système comme un ensemble d'équations non homogènes à 3 inconnues. Il est alors possible de résoudre ce système comme un problème de minimisation aux moindres carrés (avec équations normales, la méthode de la matrice pseudo-inverse ou encore la décomposition en valeurs singulières).

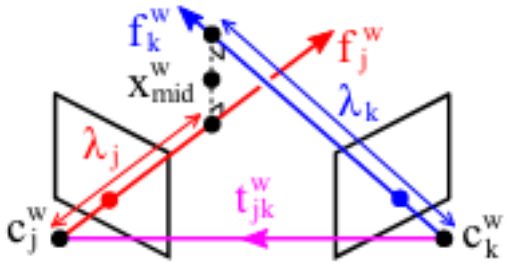


FIGURE 2.9 : Illustration de l'algorithme *midpoint* avec 2 vues. Tiré de Lee and Civera (2020)

2.6 Les systèmes optiques et autres modalités

D'autres systèmes de capture du mouvement ne se basent pas uniquement sur des images ou des vidéos. De nombreuses modalités différentes sont employées et parfois combinées avec les données visuelles issues d'images ou vidéos. Dans cette section, elles sont présentées et discutées par rapport au choix du système multivue effectué. Ces capteurs sont autant, voire plus précis, que les méthodes basées sur des images, mais possèdent des contraintes fortes vis-à-vis de leur installation ou des sujets filmés. Par exemple, ils demandent parfois d'avoir des dispositifs fixés sur les personnes (c'est le cas pour les unités de mesure inertielle ou la capture de mouvement optique avec marqueurs). Comme expliqué en 2.2 une contrainte forte du système est l'absence de dispositif intrusif pour les personnes dont la posture est extraite. Cependant, il est intéressant d'expliquer leur fonctionnement pour comparaison.

2.6.1 Capture de mouvement avec marqueurs

Le principe de la capture de mouvement (MoCap) optique avec marqueur consiste à localiser des marqueurs avec plusieurs cameras infrarouge (typiquement une dizaine de capteurs) par triangulation (voir 2.5) puis à les suivre tout au long de l'acquisition ou en post-traitement. La nature des caméras et des marqueurs varie en fonction du système. Si les marqueurs sont dits "passifs" ils sont rétro réfléchissants et renvoient une lumière infrarouge émise près des capteurs. Si ce sont des marqueurs "actifs" ils sont équipés de LED qui émettent eux même la lumière afin d'être mieux suivis (en les faisant clignoter très vite séquentiellement). Le fonctionnement standard d'un système de capture du mouvement est le suivant :

- Calibration des caméras
- Capture du mouvement
- Détection 2D des centres des marqueurs dans les images
- Reconstruction 3D des marqueurs
- Post traitements pour compléter les trajectoires manquantes

Dans les systèmes récents vient s'ajouter à ce processus une calibration du sujet afin d'appliquer des modèles d'identification automatique des marqueurs adaptés à la personne filmée. Il est souvent nécessaire de compléter des trajectoires manquantes après l'acquisition à cause des occultations, du bruit ou de mauvaises identifications des marqueurs. Ce processus se fait semi-manuellement par interpolations et filtrage des trajectoires bruitées. Les systèmes avec marqueurs actifs sont souvent plus précis et nécessitent beaucoup moins de ré-appariement des marqueurs virtuels en post traitement, mais sont aussi plus couteux et nécessitent une alimentation en énergie. Certains dispositifs combinent aussi la capture optique avec des centrales inertielles.

(a) MoCap avec marqueurs passifs (Qualisys), à l'AIHM Euromov Digital Health in Motion à Alès.

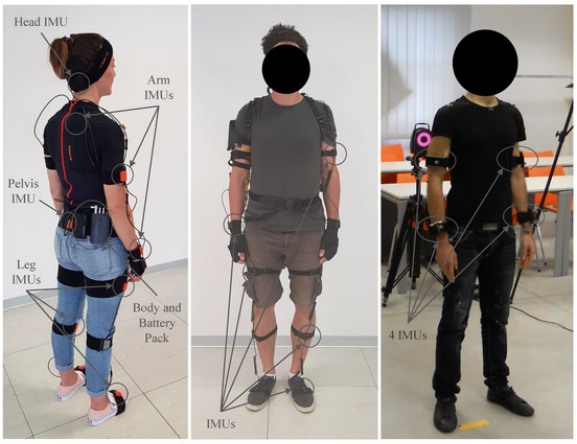
(b) MoCap active (PhaseSpace), au VR Lab de Universisty of Texas.

Les performances et le prix des systèmes utilisés sont très variables en fonction du nombre et de la nature des caméras utilisés. Pour une installation "standard" avec une dizaine de caméras, les prix varient (à la date d'écriture de ce manuscrit) de 15 000 à 200 000 euros. Avec cette installation, d'après les différents fabricants, la précision de détection des marqueurs en 3D varie de 1 à 0.001 mm.

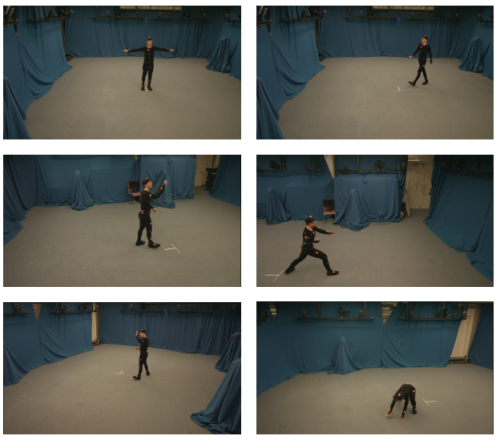
2.6.2 Centrales inertielles

La capture de mouvement inertielle (parfois désigné comme IMC pour *"Inertial Motion Capture"*) nécessite la fixation sur le sujet d'unité de mesure composée d'accéléromètres, de gyroscopes et/ou de magnétomètres. Ces capteurs sont appelés unité de mesure inertielle (IMU). Elles permettent de calculer l'accélération et la vitesse angulaire et l'orientation des parties du corps sur lesquelles elles sont placées. Un filtre de Kalman est souvent appliqué à ces différentes sources de données avant d'estimer la position de la posture.

Bien que nécessitant aussi un ensemble de capteurs placés sur le sujet, les IMU permettent de s'affranchir d'un nombre important de caméras infrarouge nécessaire dans les systèmes de capture de mouvement avec marqueurs optiques. Cependant, le prix des unités de mesure inertielles reste élevé (de 500 à 2000 euros par unité) ce qui fait augmenter très vite le coût des systèmes en fonction du nombre de personnes à suivre simultanément. L'avantage principal de cette technologie reste qu'elle n'est pas sensible aux occultations et permet un déplacement dans un large rayon autour de la centrale réceptrice du signal. En revanche, les IMU peuvent être perturbées par des interférences électromagnétiques, ce qui peut limiter leur utilisation dans un cadre industriel Slama et al. (2022). Il a également été reporté que ce type de capteur tend à dériver sur des périodes de longue utilisation Colyer et al. (2018) (défaut qui peut en partie être compensé en post traitement).



(a) Différents systèmes inertiels testés dans Giannini et al. (2020).



(b) Combinaison de multivue video, capture de mouvement avec marqueurs et IMU pour le jeux de données Total Capture (Trumble et al., 2017).

La capture de mouvement inertielle est utilisée avec succès pour l'analyse biomécanique du mouvement comme la marche (Mayagoitia et al., 2002). Il est aussi possible de l'utiliser pour l'étude de mouvements dans des environnements divers comme pour la natation (Dadashi et al. (2012), Magalhaes et al. (2015)). Cependant, certaines études montrent que les systèmes inertiels restent moins fiables que les systèmes optiques pour détecter des changements soudain de l'angle de certaines articulations (Baek et al., 2022) (ce qui permettrait d'après l'étude de détecter les tensions forte du ligament croisé par exemple).

Enfin, les IMU sont aussi utilisées en parallèles d'autres techniques pour les compléter. C'est le cas avec de la capture de mouvement optique (Lee et al., 2021) (ici un gant muni d'IMU et de marqueurs passifs plus une caméra stéréo permettent le suivi de la pose de la main et non pas l'estimation de posture du corps), mais aussi en l'alliant avec des images (Huang et al., 2019) ou vidéos (Trumble et al., 2017) multivue.

2.6.3 Caméras de profondeur

La dernière famille de capteur utilisée pour la capture du mouvement sont les caméras de profondeur (voir figure 2.10 (a)). Il en existe une grande variété et n'ont pas du tout le même fonctionnement. Nous aborderons les deux techniques suivantes :

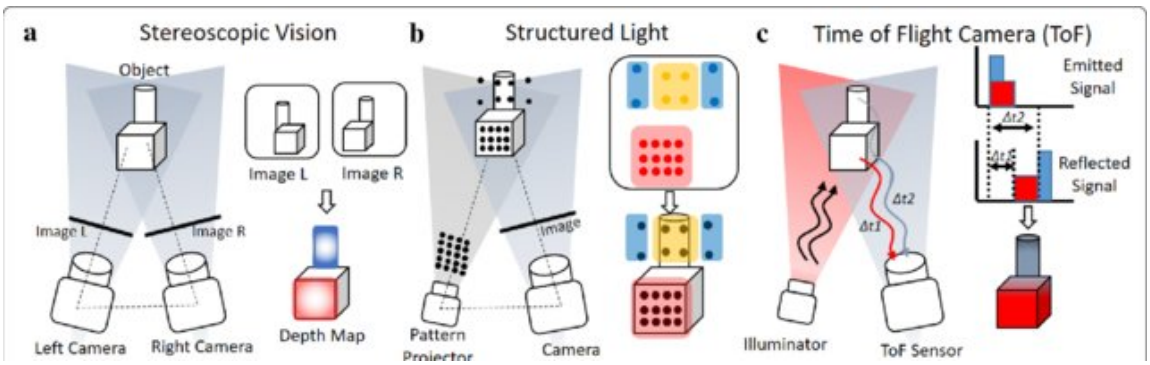
- Les capteurs actifs émettant de la lumière dans la scène
  (Plus spécifiquement les caméras à lumière structurée et caméras temps-de-vol)
- Caméras stéréo

Les caméras 3D actives produisent des images de profondeurs à partir d'un signal lumineux émis vers la scène. Le premier type de caméra active projette dans la scène un motif lumineux connu qui est capturé par une (ou plusieurs) caméras calibrées. Il est alors possible de mesurer la déformation des motifs d'une caméra à l'autre et d'en déduire la profondeur. C'est cette technologie qui est utilisée dans la première version de Kinect. Elles ont en général une précision supérieure à celle des caméras temps-de-vol (décrites plus bas) mais sont plus difficiles à calibrer et plus sensibles à la lumière naturelle.

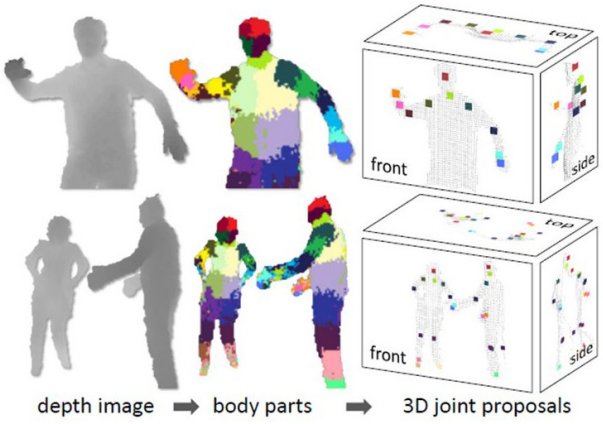
Les caméras temps-de-vol ("*time of flight*" ou TOF) captent la lumière réfléchie, ce qui permet de calculer la distance pour chaque point de la scène à l'aide du temps entre émission et réception du signal. Ce type de dispositif permet une estimation de la profondeur précise et en temps réel. Cette précision peut aller jusqu'au submillimétrique en fonction du capteur utilisé (en utilisant des lasers commeAndersen et al. (2013)) et de la distance à l'objet filmé. Cependant, les systèmes permettant de filmer sans danger des personnes capturent la profondeur avec une erreur de 1 à 5 mm pour une personne filmé entre 1 et 5 mètres. C'est le cas par exemple du capteur de profondeur du Kinect v2 (Kurillo et al., 2022).

Les caméras 3D actives permettent d'avoir une silhouette humaine précise en image de profondeur. Mais, comme il n'y pas d'identification directe des articulations, il est nécessaire d'appliquer un traitement plus important pour obtenir la posture des personnes que dans les méthodes avec marqueurs. Shotton et al. ont mis en place la méthode de reconnaissance de posture de la première version de Kinect. Ils entraînent un algorithme de forêt d'arbres décisionnels aléatoires pour obtenir une classification par pixels de l'appartenance à des zones du corps humain. Le dispositif fonctionne en temps réel et se base sur un unique capteur de profondeur en lumière structurée.

Les caméras stéréo sont des dispositifs possédant deux caméras placées à une certaine distance l'une de l'autre afin d'obtenir un système stéréoscopique comme décrit en 2.5.2. Il est alors possible d'obtenir une partie de l'information sur la profondeur de la scène filmée. En revanche, un unique capteur ne permet de pas de résoudre le problème des occultations. De plus, la précision pour l'estimation de profondeur et moins importante que pour les caméras temps-de-vol surtout pour des distances supérieur à 5 mètres. L'avantage principal sur les caméras temps-de-vol est la capacité à fonctionner dans des environnements extérieurs. Comme pour les TOF, il n'y pas de marqueurs à suivre et pour obtenir la posture de la personne dans les images de profondeur, il est nécessaire de les traiter. Lallemand et al. (2014), par exemple, entraînent également un classificateur par forêt aléatoire.



(a) Schémas du fonctionnement de trois types de caméras de profondeur. Extrait de Yang et al. (2019).



(b) Estimation de posture avec Kinect v1 dans Shotton et al.

(c) Images infrarouge et cartes de profondeur avec le système multivue avec plusieur Kinect Azure et marqueurs réfléchissant de Chatzitofis et al. (2022)

FIGURE 2.10 : Schémas et exemple illustrant les caméras de profondeur et leur utilisation pour l'estimation de posture

De manière générale, ce type de technologie fournie une information plus riches sur la géométrie des scènes et le mouvement des individus filmés, mais ne résout pas le problème majeur des occultations. Pour y remédier, il est possible d'utiliser plusieurs capteurs de profondeurs, mais plus difficilement avec les caméras actives qui vont produire des interférences dans les signaux lumineux projetés dans la scène (superposition des motifs pour la lumière structurée et réflexion de la lumière infrarouge provenant de la mauvaise caméra pour le temps-de-vol). Il existe aussi des méthodes qui combinent les capteurs de profondeurs et des marqueurs réfléchissants comme Chatzitofis et al. (2022).

2.6.4 Synthèse et limitations

Modalité des capteurs	Robustesse SMarq.	Robustesse Nat.	Prix(euros)	Précision(mm)
Capture de mouvement optique (Avec marqueurs)	✗	✗	[8k, 200k]	[0.001, 1]
Capture de mouvement inertielle	✗	✓	[6k, 30k]	[24, 46]
Caméras de profondeur				
Active	✓	✗	[400, 200k]	[20, 60]
Stéréo	✓	✓	[80, 300]	[30, 70]
Estimation de posture basée image				
Monoculaire	✓	✓	[50, 1000]	[40, 50]
Multivue	✓	✓	[300, 4000]	[20, 30]

TABLE 2.1 : Comparaison des différents types de capture du mouvement en fonction des capteurs utilisés. SMarq. : sans marqueurs, Nat. : fonctionne en lumière naturelle. La précision de la capture de mouvement optique correspond à la précision de détection des marqueurs et non pas des articulations directement.

La capture de mouvement optique constitue la technologie de référence en termes de précision, mais reste toujours coûteuse et intrusive du fait des marqueurs obligatoires pour le suivi des trajectoires des articulations. De plus, elle ne fonctionne aussi pas toujours en lumière naturelle et peut être bruité par la présence d'objets réfléchissant dans la scène. Elle nécessite aussi une étape de post traitement semi-manuelle plus ou moins importante après l'acquisition pour s'assurer que les trajectoires de marqueurs suivies sont les bonnes. La capture de mouvement avec marqueurs possède cependant les résultats les plus stables et sert de point de comparaison pour les nouvelles méthodes.

La capture de mouvement inertielle est une solution qui limite le nombre de caméras tout en résolvant le problème des occultations. Sa précision n'est pas affectée

par un environnement extérieur en lumière naturelle, mais tend parfois à dériver au cours d'une utilisation trop longue. Plusieurs méthodes d'estimation de posture avec ce type de technologie ont été développées (Trumble et al. (2017) ; Zhang et al. (2020)). Celles-ci obtiennent des résultats variants de 24 à 46 mm d'erreur moyenne, et ce, en analysant des vidéos des sujets conjointement aux données des IMUs. C'est une technologie moins coûteuse que la capture avec marqueurs classique, mais dont le prix croit vite proportionnellement au nombre de sujets dont il faut suivre le mouvement. Enfin, les IMUs reste placées sur les personnes, tout comme les marqueurs, et doivent être rechargées comme pour la capture de mouvement avec marqueurs actifs.

Les caméras de profondeurs de différentes technologies lèvent la contrainte des marqueurs, mais permettent une reconstruction 3D de la scène partielle sans résoudre le problème des occultations. Les caméras stéréo fonctionnent sous tout type de lumière, mais sont moins précises que les caméras TOF ou avec lumière structurée. Ces dernières souffrent cependant du même problème que la capture de mouvement avec marqueurs avec les objets réfléchissants et la lumière du soleil (car utilisant la lumière infrarouge pour fonctionner). Lallemand et al. (2014) reportent une erreur de localisation des articulations variant de 30 à 70 mm en utilisant une caméra stéréo pour estimer la posture.

Pour les systèmes obtenant des images de profondeur à partir de signaux lumineux (temps-de-vol ou en utilisant un motif de lumière prédéfini), les systèmes Kinect des différentes générations sont prédominants du fait de leur spécialisation initiale dans l'extraction du mouvement humain (pour le jeu vidéo d'abord, puis pour une utilisation plus médicale/industrielle ensuite). Faity et al. (2022) reportent une erreur quadratique moyenne dans le de mouvement de certains membres pour des patients post-AVC autour de 20 à 60 mm (comparativement à ce qui est obtenu à partir d'un système de capture de mouvement avec marqueur). Le prix de ce type systèmes comparativement similaire au prix d'une caméra vidéo haute définition et donc bien plus accessible que les technologies avec marqueur et inertielles. Il existe aussi des scanners 3D pour le corps humain utilisant le même type de technologie, mais pour une utilisation différente, avec un volume de capture moindre et peu ou pas de mouvements.

Comparativement à tous ces systèmes (voir tableau 2.1), l'estimation de posture sans marqueurs à partir d'images couleur possède plusieurs avantages. Elle n'utilise pas de marqueurs et fonctionne sous tout type d'éclairages (spécifications nécessaires pour le prototype), leur prix est minimal par rapport aux solutions précédentes et fournissent une précision équivalente ou légèrement plus faible (sauf par rapport aux méthodes avec marqueurs).

2.7 Conclusion

Ce chapitre expose le contexte scientifique de mes travaux de thèse et plus particulièrement les différents champs de recherche qui sont utilisés pour la reconstruction numérique du mouvement humain dans l'espace. Les méthodes d'apprentissage profond ont permis une identification de plus en plus fine de la posture humaine dans les images et les scènes en trois dimensions. Le chapitre 3 aborde ces méthodes plus en détail.

Cependant, les méthodes multivues sans marqueurs permettant une estimation de la posture 3D robuste restent rares, mais pourtant cruciales pour certaines analyses ou diagnostiques médicaux. Les travaux présentés dans le chapitre 4 explorent les caractéristiques saillantes (trajectoire de posture 2D, silhouette) pouvant être utilisées pour cette estimation 3D. Plus particulièrement, la fusion de données multivue est expérimentée afin de déterminer quelles données et traitements sont le mieux adapté à l'estimation de posture.

Le prototype dans son ensemble est décrit dans le chapitre 5. Il s'inscrit dans les méthodes d'apprentissage avec des réseaux de convolution pour une estimation 2D de la posture dans chaque vue. Il s'appuie également sur les connaissances actuelles sur les scènes 3D en vision par ordinateur décrite plus haut dans ce chapitre. Enfin, il est évalué à l'aide d'un système de capture du mouvement optique avec marqueurs.

Chapitre 3

État de l'art de l'estimation de posture sans marqueurs

Contents

3.1 Introduction

Le chapitre qui suit est tiré en partie des travaux que nous avons publiés dans Computer Vision and Image Understanding sous le titre "A review of 3D human pose estimation algorithms for markerless motion capture" (Desmarais et al., 2021).

L'estimation de la posture humaine consiste à extraire les configurations du corps dans des images ou des vidéos. Typiquement, il s'agit de l'inférence des coordonnées des articulations et de la reconstruction d'une représentation sous forme de squelette de la posture. Au cours des dernières années, l'estimation de posture 2D a dépassé le seuil 90 % d'articulations correctement détectées (Newell et al., 2016). Ces progrès ont été possibles en grande partie grâce au succès des réseaux de neurone convolutifs (CNN) et à l'apparition de jeux de données accessibles à grande échelle (Sigal et al. (2010), Ionescu et al. (2014)). Cependant, ce n'est que récemment que ces nouvelles architectures ont été déployées pour résoudre des problèmes similaires en 3D. Le défi pour ces nouvelles méthodes d'estimation de posture sans marqueur en 3D est d'être compétitives face aux techniques classiques et aux systèmes de capture de mouvement basés sur des marqueurs. L'objectif ultime serait une reconstruction 3D complète et précise du mouvement d'un individu à partir de simples images monoculaires, avec une tolérance aux occultations sévères. Cet idéal étant irréaliste, les résultats obtenus sur des tâches similaires semblent indiquer qu'il est possible d'atteindre certaines de ces conditions, même si toutes ne sont pas remplies.

D'après la méthodologie de Moeslund and Granum (2001) il est possible de comparer les systèmes de capture du mouvement selon trois principaux critères :

- L'exactitude
- La vitesse
- La robustesse

La pertinence de ces critères pour différents domaines (surveillance, contrôle ou analyse) est variable. Cependant, bien que les besoins de ces domaines d'application aient évolué, cette analyse reste pertinente. Dans ce chapitre, notre objectif principal est de guider les choix des développeurs, ingénieurs et chercheurs qui souhaitent s'appuyer sur les algorithmes étudiés. Nous comparons les méthodes académiques récemment publiées en fonction de leur pertinence pour différents domaines d'application. Tout d'abord, nous décrivons les métriques et les jeux de données de référence qui sont couramment utilisés pour l'évaluation des méthodes dans la section 3.2. Ensuite, nous détaillons et expliquons les familles d'architectures utilisées pour l'estimation de la posture humaine dans la section 3.3. Puis, nous présentons l'état actuel de l'art des techniques actuelles d'estimation de posture 3D sans marqueur en mettant l'accent sur

des articles soigneusement sélectionnés dans la section 3.4. Enfin, une analyse globale des indices de performance en matière d'exactitude, de robustesse et de vitesse est disponible en 3.5.

3.2 Évaluation des systèmes d'estimation de posture

Cette section décrit comment évaluer les méthodes d'estimation de posture sans marqueur en 3D. Pour comparer ces méthodes, nous utilisons les évaluations fournies par plusieurs jeux de données de référence. Les concepteurs de ces jeux de données recommandent souvent différentes métriques. Nous commençons par décrire comment ces métriques sont calculées puis discutons par la suite de leur pertinence dans différents contextes. Pour chacun de ces benchmarks, la quantité d'images, les environnements et les modalités d'acquisition sont également détaillés.

3.2.1 Métriques

Plusieurs métriques mesurent l'exactitude des algorithmes d'estimation de posture en 3D. Certaines calculent l'erreur moyenne, d'autres un taux de détection avec un seuil prédéfini et enfin certaines utilisent des critères perceptifs ou structurels. Dans cette sous-section, ces métriques sont décrites et leurs forces et faiblesses sont discutées dans différents contextes.

La plupart des métriques qui suivent sont utilisés dans le cadre de représentation "squelettique" (voir 2.3) du corps humain (à l'exception de celles basées sur des volumes et surfaces). Dans ce cas, il est fréquent de voir référencé dans la littérature les points d'intérêts détectés comme des "points d'articulations" (*joints*). Cependant, ces points ne correspondent pas nécessairement à des articulations au sens biologique du terme (par exemple : l'articulation de la tête ou du bassin). C'est également le cas pour les segments reliant ces points qui sont parfois appelés "membres" (*limbs*).

MPJPE (*Mean Per Joint Position Error)* : Erreur moyenne de position par articulation. Il s'agit de l'une des métriques les plus fréquemment utilisées dans la littérature. Elle est aussi parfois appelée erreur moyenne de reconstruction ou erreur 3D. La MPJPE est la moyenne des distances euclidiennes entre les coordonnées estimées et les coordonnées de la vérité du terrain pour chaque articulation. :

$$MPJPE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|m_{\hat{x}}(i) - m_x(i)\| \tag{3.1}$$

N représente le nombre d'articulations traitées, $m_{\hat{x}}(i)$ est la fonction d'estimation des coordonnées de la ième articulation et m_x est la position de vérité terrain de l'articulation. L'équation 3.1 est la mesure du MPJPE pour une image et un squelette. Pour généraliser aux vidéos, la moyenne de la MPJPE de chaque image de la séquence est calculée.

MPJPE est la métrique de base qui peut être utilisée pour évaluer une grande variété de méthodes tant qu'elles ont pour but d'estimer la position des coordonnées et la structure globale du squelette. Il peut également être adapté pour évaluer les méthodes qui n'estiment pas le même nombre de points d'intérêt que le nombre de marqueurs utilisés dans les jeux de données courants ainsi que les méthodes estimant les poses relatives au lieu de la position 3D absolue (Sigal et al., 2010). Ce type de méthode est réalisée par alignement de Procrustes (ajustement aux poses de la vérité de terrain) en utilisant un point "d'articulation" racine choisie, comme le centre du bassin. On parle parfois de N-MPJPE ou de P-MPJPE selon que l'alignement se fait uniquement par échelle ou également par rotation et translation. Les principaux inconvénients de ces mesures, identifiés par Ionescu et al. (2014), sont leur faible robustesse aux erreurs aberrantes et le fait qu'elles peuvent être influencées par des variations sans rapport avec la perception humaine de la posture (par exemple, un décalage important de plusieurs centimètres sur une coordonnée produira une erreur importante alors que la sémantique de la pose restera toujours aisément identifiable par un observateur).

MPJVE : Erreur de vélocité moyenne par articulation. Introduite par Pavllo et al. (2019), cette métrique peut être utilisée lorsqu'une séquence de pose est extraite d'une vidéo. Ici, la position absolue des articulations obtenue à partir du MPJPE est insuffisante. Pour cette raison, l'auteur utilise "le MPJPE de la première dérivée des séquences de pose 3D" pour "mesurer la régularité des prédictions dans le temps". Cette technique est utile pour comparer les modèles d'estimation qui utilisent des données temporelles.

Métriques angulaires : Une autre approche consiste à mesurer les erreurs angulaires des segments articulaires. Ionescu et al. (2014) proposent l'erreur angulaire moyenne par articulation (MPJAE) qui est aussi parfois appelée erreur angulaire moyenne (Agarwal and Triggs, 2006) :

$$MPJAE(x, \hat{x}) = \frac{1}{3N} \sum_{i=1}^{3N} |(m_{\hat{x}}(i) - m_x(i) mod \pm 180)| \tag{3.2}$$

Ici, m_x et $m_{\hat{x}}(i)$ font référence aux angles 3D pour la vérité terrain et la prédiction, respectivement. Ces métriques peuvent être utilisées lorsque l'analyse principale

est effectuée sur les angles entre deux membres spécifiques plutôt que sur le corps entier, comme pour la rééducation ou le mouvement sportif. Cependant, les auteurs de Ionescu et al. (2014) signalent que cette métrique peut avoir peu de signification perceptive. Elle peut également être difficile à interpréter, car les angles sont calculés localement et les articulations reliées à une autre articulation mal détectée peuvent ne produire aucune erreur malgré un squelette globalement désaligné. Par conséquent, ces mesures sont moins utilisées dans les publications récentes en vision par ordinateur.

Métriques de seuils : Une approche commune dans l'estimation de la posture humaine en 2D et d'autres tâches de détection consiste à définir un seuil d'erreur à partir duquel un point d'intérêt est correctement détecté. Ensuite, les statistiques des articulations correctement prédites sur un ensemble d'images peuvent être calculées. Le pourcentage de membre correct (PCP) utilise la moitié de la taille du segment de vérité terrain pour déterminer si une prédiction sur un segment de membre est correcte. Une version 3D du PCP peut également être utilisée. Un membre est correctement estimé si l'expression suivante est respectée :

$$\frac{\|\alpha - \hat{\alpha}\| + \|\omega - \hat{\omega}\|}{2} \leq \theta \|\alpha - \omega\| \tag{3.3}$$

α et ω sont les deux coordonnées mesurées des extrémités du membre et $\hat{\alpha}$ et $\hat{\omega}$ leurs prédictions. θ est un paramètre choisi pour contrôler l'exactitude requise pour le seuil (généralement 0,5). Cette métrique a été utilisée pour évaluer l'estimation de posture basée sur un modèle à l'aide de "*Pictorial Structures*" telle que Burenius et al. (2013). Le problème de cette métrique est qu'un membre plus court aura moins de chances d'être considéré comme détecté lorsque le seuil diminue.

Une autre métrique de seuil utilisée dans l'estimation de posture 2D est le pourcentage de points clés corrects (PCK). Cette métrique n'a pas le problème des membres plus courts plus difficiles à considérer comme détecté car elle utilise un seuil spécifique au sujet pour chaque articulation individuelle. Elle est calculée à partir d'une portion d'une longueur de membre fixe (par exemple 0,5 fois la longueur entre la base du cou et le haut de la tête, souvent appelée PCKh@0.5, h pour *head*). De cette façon, la métrique s'adapte automatiquement aux sujets ayant des proportions différentes, sans biais sur la taille spécifique des membres d'un individu. Mehta et al. (2016) proposent une version 3D du PCK utilisée comme principale métrique d'évaluation pour le benchmark MPI-INF-3DHP (Mehta et al., 2016). Une articulation est considérée comme détectée avec cette condition :

$$\|\alpha - \hat{\alpha}\| \leq \theta \|k - h\| \tag{3.4}$$

α et $\hat{\alpha}$ sont l'articulation cible et sa prédiction. θ est un paramètre contrôlant la

fraction de la longueur d'un "membre" (segment entre deux points "d'articulation") de référence, *k* et *h* sont les coordonnées des extrémités de ce segment (tête, torse...). Une autre solution consiste à choisir un seuil fixe arbitraire de 150 mm, ce qui fait perdre la spécificité au sujet.

L'utilisation de ces mesures basées sur des seuils est justifiée lorsqu'il s'agit de comparer des méthodes qui pourraient avoir une bonne exactitude globale, mais produire des erreurs dans un scénario spécifique (pour des articulations ou des squelettes particuliers). Cependant, cela se fait au prix d'une perte de sensibilité qui pourrait être pertinente lors de l'analyse de coordonnées locales précises (à l'échelle du millimètre), comme dans les applications biomécaniques.

Mesures basées sur des volumes ou surfaces : Certaines techniques d'estimation de la pose humaine nécessitent une mesure sur des surfaces. Ce type de métrique se retrouve dans l'estimation de pose dense Densepose (Güler et al., 2018). Cette tâche vise à récupérer la surface de l'ensemble du corps humain, et pas seulement quelques points clés articulaires. Les métriques basées sur la distance géodésique sont souvent utilisées dans ce contexte. Un exemple serait la similarité de points géodésiques décrite dans Güler et al. (2018) :

$$GPS(j) = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2k^2}\right) \tag{3.5}$$

P_j est un ensemble de points représentant la surface du corps de la jème personne. $g(i_p, \hat{i}_p)$ est la distance géodésique calculée entre le point estimé et celui de la vérité terrain (obtenue à partir d'annotations manuelles). Un score GPS de 0,5 indique que cette distance est égale à la moitié d'une distance prédéfinie réglable avec le paramètre k (souvent défini comme une fraction d'un segment d'articulation).

"Average precision" ou AP est la métrique utilisée dans le jeu de donnée MS-COCO (Lin et al., 2015) sur lequel est basé Densepose. Elle est utilisée pour la détection d'objet, mais aussi pour la segmentation sémantique. Dans ces cas, cette métrique mesure la similarité entre les objets prédits et les vérités terrains (respectivement les boîtes englobantes et les masque de segmentation). Par exemple, pour la détection d'objet, cette similarité est calculée à partir de "l'IoU" (*intersection over union*) de ces deux boîtes englobantes. Le calcul de l'AP est ensuite l'intégrale de la courbe précision-rappel. À partir de 2016, MS-COCO propose une labellisation par points d'intérêt de la posture des personnes dans les images qui en contiennent. Ainsi une adaptation de la métrique de l'AP pour la tâche détection de ces points d'intérêts est proposée : pour mesurer la similarité entre l'ensemble de points d'articulation détectés et les vérités terrains l'OKS

(*"l'object keypoint similarity"*) remplace l'IoU :

$$OKS = \sum_i \frac{\exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)} \tag{3.6}$$

d_i correspond à la distance euclidienne entre le ième point détecté et sa vérité de terrain. v_i est un drapeau de visibilité pour ne pas prendre en compte les points non visible dans l'image. Enfin, s et k contrôle l'échelle de l'objet. Un score de similarité compris entre 0 et 1 est obtenu pour chaque point et l'AP peut ensuite être calculée de la même manière que pour les IoU. Les méthodes d'estimation de posture qui sont évaluées sur le jeu de donnée MS-COCO utilisent couramment cette métrique.

Il existe aussi des métriques pour le suivi de forme humaine en 3D est une autre variante de la tâche consistant à reconstruire et à suivre le volume du corps humain image par image dans une vidéo. Une approche courante utilise l'algorithme itératif du point le plus proche (ICP) pour ajuster les données d'image à un modèle. Huang et al. (2018) suggèrent d'utiliser des forêts aléatoires et un appariement aux plus proches voisins avec deux caractéristiques volumétriques basées sur les voxels et la tessellation de Voronoï centroïdale au lieu de l'ICP.

Enfin, une autre famille populaire de méthodes utilise des données multivues et des techniques de "*shape-from-silhouette*" pour créer des représentations volumétriques du corps humain. Ce volume peut ensuite être utilisé pour la prédiction de la localisation des articulations. Ces méthodes produisent des "*Probabilistic Visual Hulls*" (PVH) (Grauman et al., 2003) dans des grilles de voxels. Bien que l'estimation de la forme du corps humain ne soit pas l'estimation de la pose humaine, il s'agit d'une tâche proche qui peut être utilisée à différentes étapes d'un système modulaire de capture de mouvement. Avec des ensembles de données multivues, il est facile d'obtenir des PVH vérité de terrain qui peuvent ensuite être utilisées pour évaluer les volumes reconstruits en 3D (par exemple Trumble et al. (2017)). Pour ces méthodes, l'erreur quadratique moyenne peut être calculée à partir de la grille de voxels.

Chacune de ces mesures peut être utilisée dans des variations spécifiques ou des cas limites de la tâche. Pour l'estimation "classique" de la pose humaine, la MPJPE semble plus populaire, car elle est simple et aucun paramètre supplémentaire n'intervient dans son calcul. Cependant, certains articles publiés (Ionescu et al. (2014), Mehta et al. (2016)) affirment que les métriques à seuil sont plus efficaces pour identifier les erreurs sur des articulations spécifiques et moins enclines à pénaliser les erreurs non pertinentes sur le plan perceptif. Enfin, pour évaluer les méthodes qui traitent des vidéos et produisent des séquences de pose 3D, le MPJVE est une bonne alternative pour mettre en avant les techniques qui produisent des mouvements humains plus réalistes.

En outre, les métriques décrites ci-dessus n'expriment que l'exactitude physique de plusieurs manières, les métriques basées sur des seuils introduisant parfois des paramètres perceptifs. Cependant, selon le cas d'utilisation, il peut être pertinent de prendre en considération des métriques perceptuelles plus complexes (Marinoiu et al. (2016), Marinoiu et al. (2013)) ou des métriques structurelles (Kocabas et al., 2019b). Elles peuvent être utiles lorsque des informations de pure position produisent le même score d'erreur pour deux poses prédites différentes. Des travaux importants ont été réalisés sur la façon dont les humains perçoivent ce qu'est une configuration valide et réaliste du corps humain. Ces métriques peuvent être utiles dans des domaines qui ne sont pas concernés par les contraintes biologiques et physiques, mais davantage par la sémantique de la pose.

3.2.2 Jeux de données de référence

La collecte de données précises pour l'estimation de la pose humaine est un processus long et complexe qui dépend des progrès des technologies d'acquisition. De plus, plusieurs choix spécifiques sont nécessaires concernant la modalité du capteur, sa quantité et le protocole d'acquisition.

La complexité de cette tâche explique pourquoi il a fallu du temps à la communauté scientifique pour créer des jeux de données de références contenant une grande quantité d'images : aujourd'hui, il existe de nombreuses variations entre les environnements monoculaires et multivues, les environnements contrôlés en laboratoire et les environnements en condition réelle ("*in-the-wild*") (voir tableau 3.1), etc. Avec les nouvelles solutions commerciales qui commencent à produire des résultats similaires à ceux de la capture de mouvement traditionnelle, certains jeux de données commencent également à utiliser des vérités-terrain labellisées sans marqueurs (par exemple, Kanko et al. (2020) avec le système Theia). Elles ont l'avantage de fournir facilement des images dans tout type d'environnement. Cependant, l'utilisation de ce type de données comme vérité terrain peut être remise en question, car elles sont elles-mêmes obtenues par des méthodes qui ne sont pas toujours disponibles et transparentes. Malgré cette diversité, il n'existe encore que quelques références académiques librement accessibles contenant plus d'un million d'images.

Les jeux de données jouent un rôle important car ils sont utilisés pour valider et tester les algorithmes, mais aussi pour entraîner et "*fine-tune*" les modèles d'apprentissage profond. Fournir d'importants jeux de données de haute qualité avec d'excellents labels et une grande variété de poses est un défi majeur. Trois références répondent historiquement le mieux à ces critères : HumanEva I et II (Sigal et al., 2010), Human3.6M (Ionescu et al., 2014) et Total Capture (Trumble et al., 2017). Elles contiennent des sé-

Jeu de donnée	#Image #Séquence Vidéo	#Sujet	#Vue	Résolution	Fréquence	Profondeur	IMU	Environnement de capture et caractéristiques principales
Human3.6M (Ionescu et al., 2014)	3.6 millions 1 376	11	4	1000x1000	50Hz	Yes	No	Envt. de labo.
HumanEva (Sigal et al., 2010)	80 000 56	4	7	660x500	60Hz	No	No	Envt. de labo.
Total Capture (Trumble et al., 2017)	1.9 millions N/A	5	8	1920x1080	60Hz	No	Yes	Envt. de labo. Données IMU.
MPI-INF-3DHP (Mehta et al., 2016)	1.3 millions 64	8	14	N/A	N/A	No	No	"In the wild" & envt. de labo., exterieur/interieur avec écrans verts. Vérités terrain sans-marqueurs.
MuPoTS-3D (2018) (Mehta et al., 2018)	8 000 20	8	1	2048x2048 1920x1080	30Hz - 60Hz	No	No	Multipersonne, "In the wild". Scènes en exterieur et interieur. Vérités terrain sans-marqueurs.
3DPW (von Marcard et al., 2018)	> 50000 60	7	1	N/A	30Hz	No	Yes	"In the wild" en exterieur camera unique en mouvement & IMU Jusqu'à 2 sujets.
CMU Mocap (CMU)	N/A 2 605	109	1	352x240	30Hz	No	No	Envt. interieur. Divers actions et sujets.
CMU-MMAC (De la Torre et al., 2008)	≈450 000 N/A	25	5	1024x768 (x3) 640x480 (x2)	30Hz (x3) 60Hz (x2)	No	Yes	Envt. de labo. Sujets cuisinant 5 recettes.
TNT15 (von Marcard et al., 2016)	13 000 20	4	8	800x600	50Hz	No	Yes	Envt. de bureau. Pas marqueurs, uniquement données IMU.
AMASS (Mahmood et al., 2019)	N/A (>40 hours)	346	variable	variable	variable	No	No	Paramétrisation unifiée de 15 jeux de données. Modèles de maillage du corps humain.
MoVi (Ghorbani et al., 2020)	N/A (17 hours)	90	4	800x600 1920x1080	30Hz	No	Yes	Données MoCap, vidéo, de forme et IMU synchronisées.

TABLE 3.1 : Jeux de données populaires utilisés pour comparer, entraîner et tester les modèles d'estimation de la posture humaine. Les images vidéo, le nombre de sujets et d'actions donnent une indication de la diversité du jeu de données et du nombre de configurations de posture. Le nombre de vues des caméras RVB, la résolution et la fréquence d'acquisition des caméras évaluent la qualité et la quantité d'informations vidéo exploitables. Les unités de mesure inertielle (IMU) sont parfois utilisées pour affiner les résultats de la capture de mouvement ou de la détection d'une seule image. Si ce n'est pas spécifié, la méthode de capture de mouvement est basée sur des marqueurs réfléchissants.

quences vidéo avec plusieurs angles de vue associés à des coordonnées articulaires capturées par motion capture. Toutefois, ces jeux de données de grande échelle sont principalement capturées dans des environnements de laboratoire contrôlés avec des systèmes à base de marqueurs, donc avec une variété limitée d'arrière-plans, de postures et de sujets.

D'autres jeux de données se concentrent sur la diversité des poses et des sujets, ou sur l'acquisition en milieu naturel. Bien qu'intéressant pour expérimenter sur l'estimation de la posture humaine ou ses sous-tâches, ils ne fournissent pas une quantité et variation suffisante pour réaliser des évaluations à grande échelle. Toutefois, une exception notable est MPI-INF-3DHP (Mehta et al., 2016), qui est de plus en plus utilisé pour évaluer les algorithmes car il propose une grande variété de contextes et de sujets (en faisant de l'acquisition sur fond vert et en extérieur) avec une quantité importante de données.

La capture de mouvement basée sur des marqueurs est le moyen le plus facile d'obtenir quelque chose qui s'approche d'une donnée de vérité terrain. Cependant, les anciens jeux de données ont une faible résolution d'image et certains présentent des annotations inexactes pour certains sujets (rapporté par Iskakov et al. (2019)). De plus, Colyer et al. (2018) ont noté des erreurs d'environ 10mm ou 10° par rapport à des tests réalisés avec des méthodes intrusives plus proches de l'anatomie humaine réelle (ie : broches osseuses intra-corticales). Cela est dû au fait que les points d'intérêt sont reconstruits à partir de groupes de marqueurs placés sur la peau ou les vêtements des sujets. Donc des surfaces non rigides qui peuvent se déplacer légèrement au cours de l'acquisition.

Certains jeux de données proposent des modèles 3D du corps humain basés sur des maillages (Mahmood et al. (2019), Ghorbani et al. (2020)). Ces représentations peuvent être utilisées comme cibles d'apprentissage pour les algorithmes d'estimation qui se soucient d'informations plus riches que la représentation squelettique des poses et des positions articulaires. Elles sont également utiles pour les applications qui nécessitent des modèles entièrement articulable, comme l'animation. Le jeu de données AMASS (Mahmood et al., 2019) unifie plusieurs jeux de données de capture de mouvement en fournissant la représentation 3D des sujets. Pour ce faire, les poses du corps sont calculées à l'aide du modèle SMPL (Loper et al., 2015) et de sa méthode de régression (MoSh++). En addition, Ghorbani et al. (2020) fournit un nouveau jeu de données avec des données de capture de mouvement (MoCap) et de centrale inertielle, qui viennent enrichir AMASS.

3.3 Taxonomie de l'estimation de posture 3D

Dans cette section, les principales familles de méthodes d'estimation de la posture 3D seront décrites. Elles peuvent être classées comme des méthodes utilisant des modèles de corps humain, des algorithmes d'apprentissage ou des informations géométriques. Dans le cas d'une méthode d'apprentissage par réseau de neurones, des architectures de base (*"backbone"*) sont utilisées comme briques élémentaires. De plus, de nouvelles fonctions de coût adaptées aux données en entrées et sorties des réseaux sont créés. Le tableau 3.2 est la taxonomie complète de toutes les méthodes abordées dans cet état de l'art, détaillant les architectures de réseau et fonctions de coût dans chaque cas (3.3.3).

3.3.1 Modèle du corps humain

Historiquement, les algorithmes d'estimation de la posture s'appuyaient sur des modèles du corps humain basés sur des membres ou des représentations en **squelettes**. Chaque nœud représentait une articulation et les arêtes la longueur et l'orientation des membres. Un exemple de ce type d'approche est le "*Pictorial Structure Model*" (PSM) introduit par Fischler and Elschlager (1973) et utilisé pour l'estimation de la posture dans Felzenszwalb and Huttenlocher (2005), Marinoiu et al. (2013) et Belagiannis et al. (2014). À l'origine, ce modèle est conçu comme la minimisation d'une fonction d'énergie. La fonction de coût combine un terme d'erreur pour l'erreur de localisation des articulations et une pénalité pour les déformations de la longueur des segments.

D'autres méthodes adoptent un squelette humain basé sur la **cinématique**, où chaque paire d'articulations liées est représentée comme un vecteur. Des contraintes angulaires et de longueur peuvent être appliquées pour détecter les postures. Le "*Kinematic Chain Space*" proposé par Wandt and Rosenhahn (2019) en est un exemple.

Les modèles basés sur des **maillages 3D** consistent en une reconstruction de la surface du corps humain. Ces modèles offrent des informations plus riches que les modèles basés sur des "squelettes" et peuvent être utilisés pour déduire la représentation spatiale d'un sujet dans une scène virtuelle ou pour convertir une posture capturée en maillage complet pour l'animation. Un exemple de modèle de maillage humain est le modèle SMPL (Loper et al., 2015). Ce modèle est appris à partir de nombreux scans numériques de corps humains en 3D et peut être adapté à un ensemble de paramètres de posture et de forme produits par des algorithmes d'estimation de la pose.

			Modèles de corps humains			Réseaux de neurones					
Méthode	Representations intermédiaires	Fonction de coût	Chaînes Cinématiques	Squelette (ex. PSM)	Maillage (ex. SMPL)	Architecture	GAN Adv.lea.	RNN LSTM	TCN	Attention	GCNN
Pavlakos et al. (2017a)	cartes de chaleur 2D	L2				SHNet					
Mehta et al. (2017)	cartes de chaleur 2D, "Location maps"	L2	•			Resnet50					
Zhou et al. (2017)	cartes de chaleur 2D	L2, "geometric loss"				SHNet					
Martinez et al. (2017)	coordonnées 2D de la posture	L2				SHNet (2D) + MLP					
Sun et al. (2018)	cartes de chaleur 2D/3D	n'importe quelle fonction de coût applicable à une carte de chaleur.				Resnet et SHNet testés					
Omran et al. (2018)	carte de segmentation des membres	"3D and 2D joint loss" (L2) "3D latent parameter loss" (L1)			•	RefineNet (Lin et al., 2016) + Resnet50					
Mehta et al. (2018)	"Occlusion Robust Pose Maps" "Part affinity fields"	L2	•			ResNet50					
Kolotouros et al. (2019)	N/A	L2			•	ResNet50					
Wandt and Rosenhahn (2019)	N/A	"Reprojection loss" Wasserstein	•			SHNet (2D)	•				
Xu et al. (2019)	"pixel-to-surface maps"	L1 et L2			•	Resnet50	•				
Kocabas et al. (2019b)	coordonnées 2D de la posture	smooth L1				Resnet50					
Mathis et al. (2018)	N/A	L2				Resnet50 Resnet101 testé					
Mehta et al. (2020)	"3D pose encoding" "Part affinity fields"	smooth L1	•			"SelecSLS Net" Fully connected					
Hossain and Little (2018)	coordonnées 2D de la posture sequence	L2, "derivative loss" sur un ensemble d'articulations				SHNet (2D)		•			
Cai et al. (2019)	coordonnées 2D de la posture, graphes spatio-temporels	L2, "symmetry loss", "derivative loss" sur un ensemble d'articulations				CPN					•
Pavllo et al. (2019)	coordonnées 2D de la posture	MPJPE, "bone length L2 loss", "2D back-projection loss"				SHNet, CPN et Mask-RCNN testés			•		
Cheng et al. (2019)	coordonnées 2D de la posture	L2, "2D projection loss"		•		SHNet (2D)	•		•		
Cheng et al. (2020)	cartes de chaleur 2D	L2, "Multiview loss" "2D projection loss"	•			HRNet (2D)	•		•		
Liu et al. (2020)	N/A	N/A				SHNet et CPN testé (2D)			•	•	
Wang et al. (2020)	coordonnées 2D de la posture, graphes spatio-temporels	L2, "motion loss"				CPN et HRNet testé (2D)					•
Qiu et al. (2019)	cartes de chaleur 2D	L2		•		SimpleNet					
Iskakov et al. (2019)	cartes de chaleur 2D	soft L2, L1				SimpleNet					
He et al. (2020)	coordonnées 2D de la posture	L2				SimpleNet				•	
von Marcard et al. (2016)	silhouettes multivues orientations IMU	N/A	•		•	N/A					
Trumble et al. (2017)	PVH, orientations IMU puis coordonnées 2D de la posture	L2				Classical 3D CNN		•			
von Marcard et al. (2018)	coordonnées 2D de la posture, orientations IMU	N/A	•		•	Cao et al. (2017) (multipersonne)					
Huang et al. (2019)	"multi-channel volumes"	L2				SHNet (3D Conv)					
Zhang et al. (2020)	cartes de chaleur 2D	N/A		•		SimpleNet					

TABLE 3.2 : La taxonomie des méthodes examinées. En cas de multiples étapes, nous indiquons les représentations intermédiaires. Pour les méthodes d'apprentissage, les fonctions de coût et l'architecture de base sont indiquées lorsqu'elles sont présentes (pour de nombreuses méthodes à deux étapes, ces structures de base ne concernent que la première étape de la détection 2D). Ces architectures de base sont abréviées : *SHNet* : Stacked Hourglass (Newell et al., 2016), *CPN* : Cascaded Pyramid (Chen et al., 2018), *HRNet* : High Resolution Network (Sun et al., 2019)) et *SimpleNet* : Simple Baselines 2D (Xiao et al., 2018).

3.3.2 Modèles géométriques

Lorsque plusieurs caméras sont disponibles, la géométrie multivues est fréquemment utilisée pour l'estimation de la posture 3D. Une des façons de déduire les coordonnées des articulations en trois dimensions est d'utiliser la **triangulation** des coordonnées dans la scène à partir des coordonnées 2D des images de chaque vue. Selon la calibration et la disponibilité des paramètres extrinsèques et intrinsèques des caméras, différents schémas de reconstruction sont possibles.

Une autre approche consiste à **fusionner les caractéristiques** de différentes vues le long de lignes épipolaires avant d'inférer les postures. La ligne épipolaire est l'image dans une caméra d'un rayon passant par le centre optique de l'autre caméra et un point dans la scène. Si l'on considère un point dans la première vue, il est garanti que le point correspondant dans la deuxième vue se trouve sur la ligne épipolaire dans l'autre image. Grâce à ces informations préalables sur les différentes vues, il est possible d'affiner la posture 2D ou de fusionner des caractéristiques multivues avant d'estimer la posture 3D elle-même.

Enfin, d'autres méthodes utilisent une reconstruction **shape-from-silhouette** pour obtenir la forme 3D entière du corps avant la détection des articulations. Ces techniques segmentent d'abord la surface du corps des personnes dans chaque image, puis reconstruisent leur volume (un exemple de ces méthodes sont les "*probabilistic visual hull*" de Grauman et al. (2003)). Ces volumes peuvent ensuite être utilisés comme caractéristiques intermédiaires pour des méthodes d'apprentissage ou pour l'adaptation à des modèles 3D du corps humain décris plus haut.

3.3.3 Réseaux de convolution pour l'estimation de posture

Dans cette sous-section sont décrites les architectures de réseaux de neurones les plus fréquemment trouvées dans l'état de l'art pour l'estimation de posture 2D et 3D.

Avant de passer en revue les algorithmes les plus récents, il est important de considérer les architectures de base qui sont couramment utilisées pour l'estimation de la posture en 2D ou en 3D. Elles sont largement utilisées dans les approches "top-down", avant tout calcul, réduisant le problème à une mise en correspondance de coordonnées 2D et 3D. Ces approches raisonnent à partir des coordonnées articulaires de bas niveau pour déduire des informations de haut niveau sur le squelette humain (voir Fig. 3.2). À l'inverse, les techniques "bottom-up" extraient des caractéristiques des images et tentent de raisonner sur des modèles de corps humain qu'elles font correspondre aux données. Ces architectures de base sont aussi employées directement pour l'estimation de la posture en 3D. Par exemple, Huang et al. (2019) utilisent des "Hourglass Networks" (Newell et al., 2016) incluant une convolution 3D sur des représentations volumétriques. Pour le détail du fonctionnement de ces architectures de base, se réfé-

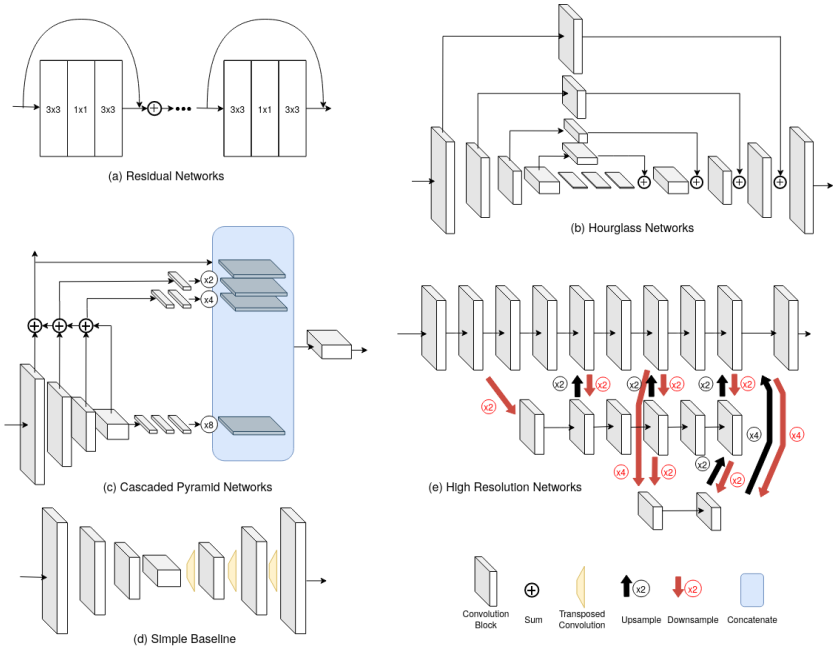


FIGURE 3.1 : Architectures de base utilisées pour l'estimation de la posture en 2D. (a) Les blocs résiduels sont la principale caractéristique des variantes de ResNet (Resnet101, Resnet50, Resnet152) (He et al., 2015). Ils sont également présents dans des modèles plus spécialisés dans l'estimation de la posture humaine : (b) réseaux "Stacked Hourglass" (Newell et al., 2016) et (c) "Cascaded Pyramid Networks" (Chen et al., 2018). Le réseau "Simple Baseline 2D" (d) proposé par (Xiao et al., 2018) utilise des convolutions transposées pour retrouver la résolution d'entrée. (e) Le réseau "élevéer Resolution Network" (HRNet) (Sun et al., 2019) traite les caractéristiques à haute et basse résolution en parallèle au sein de sous-réseaux qui partagent l'information.

rer à Xiao et al. (2018), Newell et al. (2016) ou Chen et al. (2018). Dans le présent document, nous nous limitons aux principes généraux des architectures de base couramment utilisées et nous donnons un aperçu de leur structure et de leurs performances.

Deux architectures couramment utilisées sont les réseaux en sablier ("Hourglass Network") (Newell et al., 2016) et les réseaux en pyramide en cascade ("Cascaded Pyramid Network") (Chen et al., 2018) (Fig. 3.1, (b) et (c)). Ils calculent tous deux des caractéristiques d'image à différents niveaux de résolution avec l'idée clé d'englober des caractéristiques discriminantes globales et locales. Les "Hourglass Network" sont composés de modules en sablier empilés : la première partie du module réduit la résolution au fur et à mesure qu'elle traverse les couches convolutives. La seconde partie échantillonne à la hausse les caractéristiques tout en les additionnant aux caractéristiques correspondantes de même dimension de l'étape précédente. Une supervision

intermédiaire est effectuée à la fin de chaque module. Les réseaux pyramidaux en cascade sont une architecture en deux étapes qui prédit les postures à partir d'un réseau pyramidal de caractéristiques. Elle affine ensuite les résultats pour les points difficiles à prédire. Le réseau pyramidal fusionne les caractéristiques à différentes résolutions pour produire des cartes de chaleur des positions articulaires. Dans la deuxième étape, le processus de raffinement utilise des caractéristiques intermédiaires de la pyramide à différents niveaux. Elles sont échantillonnées et additionnées avant de passer par un bloc convolutif final. Ce processus est effectué uniquement pour les "points clés difficiles à prédire" (choisis avec la fonction de coût du réseau pyramidal). Ces deux réseaux utilisent les modules de connexion résiduels de He et al. (2015) comme blocs de base de leur architecture. Basée sur la connexion par saut entre les couches convolutives, cette technique est largement utilisée dans de nombreuses tâches de vision par ordinateur pour l'extraction de caractéristiques. Certains systèmes d'estimation de posture en 2D utilisent directement l'une des variantes originales de ce réseau (par exemple, Resnet50, Resnet101, etc.).

"High Resolution Networks" de Sun et al. (2019) sont aussi des réseaux de neurones exploitant les différentes résolutions d'une même image. Dans ce cas, les résolutions sont traitées en parallèle, les couches de convolution partageant les poids par le biais de "blocs d'échange".

Le détecteur multipersonnes 2D de Cao et al. (2017) est parfois utilisé. Il utilise des "part affinity fields" qui sont "une représentation non paramétrique de la relation entre les parties du corps". À partir de ces caractéristiques et des cartes de confiance de localisation des articulations, les postures humaines sont prédites avec des associations correctes pour plusieurs sujets simultanément. Un autre avantage de cette méthode est qu'elle fonctionne en temps réel.

Enfin, Xiao et al. (2018) propose une méthode de référence utilisant également Resnet comme base. Cette technique utilise des couches de déconvolution pour produire des cartes de chaleur à partir de caractéristiques d'images profondes, sans procédure spécifique pour les points difficiles à prédire. Malgré sa simplicité, ce modèle permet d'obtenir des résultats compétitifs à un coût de calcul efficace. Il est donc utilisé pour des évaluations comparatives mais aussi parfois comme détecteur 2D de base.

3.4 Comparaison des méthodes

Cette section décrit et compare les méthodes les plus performantes pour l'estimation de la posture 3D sans marqueur basée sur la vision (voir tableaux 3.3, 3.4 et 3.5). Le processus de sélection a été le suivant :

- Les meilleures méthodes de l'état de l'art sur chacun des jeux de données de référence les plus populaires.
- Méthodes les plus citées en vision par ordinateur et dans les domaines d'application de l'estimation de la posture humaine en 3D (biomécanique, robotique, sciences du sport, interaction homme-machine, etc.)
- Méthodes récentes présentant des résultats intéressants ou des approches originales susceptibles de faire progresser la recherche.

La première observation concernant l'évolution de l'exactitude dans l'état de l'art est qu'elle s'améliore rapidement et régulièrement. Cette tâche a retenu l'attention des chercheurs et des grandes entreprises du numérique et, chaque année, de nouveaux records sont atteints. L'erreur moyenne est passée d'environ 100mm à moins de 20mm en 10 ans. Aujourd'hui un plateau semble avoir été atteint autour de ces 20mm d'erreur moyenne. Il reste à déterminer si l'exactitude des méthodes va dépasser ce plateau à l'avenir. Cependant, étant donné la grande diversité des approches et des modalités, on peut affirmer qu'à ce jour, il n'existe pas de consensus sur la meilleure méthode à utiliser. Un aperçu général des différentes approches est présenté avec la figure 3.2

Dans cette section, nous nous concentrons sur les techniques sans marqueur les plus performantes actuellement disponibles, indépendamment des capteurs ou des algorithmes qu'elles utilisent. Les méthodes utilisant des images monoculaires et des séquences vidéos sont d'abord présentées. Ensuite, nous présentons les méthodes utilisant des vues multiples. Enfin, nous incluons les méthodes utilisant d'autres modalités comme étape de pré-traitement ou pendant la prédiction, principalement les unités de mesure inertielles. Une analyse globale des indices de performance des différentes familles de méthodes sera effectuée dans la section suivante (voir 3.5).

3.4.1 Image monoculaire

Pavlakos et al. (2017a) présentent l'une des premières méthodes à proposer un réseau neuronal convolutif de bout en bout en une seule étape pour prédire la posture humaine en 3D à partir d'une seule image RVB. Pour ce faire, ils se concentrent sur la nature 3D de la tâche. Leur architecture utilise des modules en sablier empilés Newell et al. (2016) (voir architectures 2D 3.3.3) qui produit des volumes de probabilité de voxels pour chaque articulation dans un espace 3D discrétisé autour de la cible. Ils proposent également une nouvelle méthode de supervision intermédiaire inspirée des succès obtenus dans l'estimation de la posture humaine en 2D. Cette méthode originale n'utilise pas d'étape d'estimation des articulations en 2D. Au lieu de cela, elle utilise une approche "coarse-to-fine" qui exploite des cartes de chaleur 2D comme vérités terrain pour la supervision intermédiaire, puis les fusionne avec les caractéristiques de l'image comme sortie des modules suivants. Plus loin, le réseau reconstruit

Méthode	Human3.6M	MPI-INF-3DHP	HumanEva
Pavlakos et al. (2017a)	71.90	-	**24.3**
Mehta et al. (2017)	80.5*	76.6	-
Zhou et al. (2017)	64.9	69.2	-
Martinez et al. (2017)	62.9/47.7*	-	24.6
Sun et al. (2018)	64.1/49.6+/40.6*+	-	-
Omran et al. (2018)	59.9*	-	64.0*
Mehta et al. (2018)	69.9	74.1	-
Kolotouros et al. (2019)	**41.1**	76.4	-
Wandt and Rosenhahn (2019)	50.9/38.2*	82.5	-
Xu et al. (2019)	82.4/53.9*/48.0*+	73.1/76.9+/89.0*+	-
Kocabas et al. (2019b)	51.83+/45.04*+	77.5	-
Mathis et al. (2018)	-	-	-
Mehta et al. (2020)	63.6	**82.8**	-

TABLE 3.3 : Comparaison de l'exactitude de plusieurs méthodes monoculaires de l'état de l'art. Les résultats de Human3.6M et HumanEva sont rapportés en MPJPE absolue (la plus basse est la plus précise) les résultats de MPI-INF-3DHP sont rapportés en 3DPCK (la plus haute est la plus précise). Les techniques avec l'annotation + utilisent des données d'entraînement supplémentaires pour obtenir le résultat ; les autres utilisent les protocoles recommandés par le jeu de données de référence. L'annotation * indique les résultats publiés avec alignement par procrustes sur les postures de vérité terrain avant évaluation.

des cartes de voxels en 3D. La supervision est effectuée à l'aide de gaussiennes 3D autour des vérités terrain des coordonnées 3D données. Ce réseau profond fournit des résultats précis (72 mm) pour une méthode monoculaire utilisant une représentation de données purement 3D. Cependant, des méthodes plus récentes montrant de meilleurs résultats utilisent des architectures top-down à deux étapes comprenant la prédiction 2D comme première étape pour la détection 3D. Cela suggère que les caractéristiques de l'image, même transformée en caractéristiques intermédiaires adaptées, ne sont pas assez riches pour des inférences 3D directes.

VNect de Mehta et al. (2017) est un système d'estimation de la posture humaine en 3D et en temps réel. Il consiste en une architecture similaire à celle de Resnet50 qui génère des cartes de chaleur en 2D pour chaque articulation ainsi que des "cartes de localisation" nouvellement introduites. Ces cartes de localisation prédisent les positions relatives X, Y et Z d'une articulation par rapport à son articulation racine. Pour chaque articulation, l'emplacement est traité à partir du pic des cartes de chaleur 2D et la coordonnée relative à la racine est lue dans les cartes de localisation. Ensuite, un modèle

cinématique du squelette humain est ajusté aux postures prédites, afin d'améliorer la cohérence temporelle et de réduire la distorsion. Les points forts de cette méthode sont qu'elle fonctionne en temps réel et peut être utilisée dans différents environnements (notamment en extérieur). Cependant, les auteurs énumèrent plusieurs limites à leur approche, principalement le fait que les erreurs d'estimation de la profondeur peuvent conduire à des prédictions 3D erronées. La méthode n'est également capable que d'estimer la posture relative et nécessite donc une détection précise d'une articulation racine (au niveau du bassin).

Mehta et al. (2018) proposent après Vnect, une technique utilisant également des cartes de localisation, mais les modifie pour en faire des "Occlusion-Robust Pose-Maps" (ORPM) qui infèrent la posture 3D de plusieurs sujets. Ces ORPM sont similaires aux cartes de localisation mais contiennent également des informations structurelles sur la posture. Pour chaque articulation, l'emplacement dans les cartes 3D est stocké à la position de l'articulation mais aussi le long d'un ensemble prédéfini d'articulations (divisant le corps en deux bras et deux jambes) et d'articulations racines (bassin et cou). Ils utilisent également des "part affinity fields" (Cao et al., 2017), qui sont souvent employés dans l'estimation de la posture multipersonnes en 2D. Ils représentent des "champs vectoriels 2D pointant d'une articulation [...] vers ses parents". Ainsi, à partir d'articulations racines valides, toutes les articulations peuvent être déduites en suivant la chaîne cinématique du corps humain. La particularité de cette méthode est qu'elle permet de détecter plusieurs personnes en une seule fois, sans avoir recours à des estimateurs de posture 2D ou à des détecteurs de personnes standard. Elle introduit également deux nouveaux ensembles de données pour l'entraînement et l'évaluation : MuCo-3DHP (jeu de données multipersonnes à grande échelle avec occlusion, composé d'images MPI-INF-3D) et MuPoTS-3D (jeu de données multipersonnes en conditions réelles, filmé dans divers environnements avec des vérités terrain sans marqueur).

Dans Mehta et al. (2020) le système Xnect est présenté dans la continuité des deux méthodes précédentes. Il s'agit d'une méthode en trois étapes qui combine un réseau convolutif pour l'extraction de caractéristiques, une couche totalement connecté pour l'estimation de la posture et un ajustement des résultats précédents à un modèle cinématique pour affiner la cohérence de la posture. La première étape consiste à extraire les articulations 2D et leur association à un sujet par le biais de "part affinity fields" (similaire à Cao et al. (2017)) ainsi que l'encodage de la posture 3D de la même manière que les méthodes précédentes. La principale différence avec la méthode de Mehta et al. (2018) est que chaque articulation n'encode que les informations relatives à sa position par rapport à l'articulation parente et à la position de ses enfants. Pour cette étape, un nouveau module de base est présenté pour les architectures de réseau afin de réduire les coûts de calcul : SelecSLS. Il consiste en des convolutions successives 3 par 3 et 1

par 1 avec des sauts de connexion inter- et intra-module. Il atteint une exactitude similaire à celle de Resnet50 tout en étant 1,4 plus rapide à l'inférence. La deuxième étape consiste en la détection de la posture en 3D, pour chaque sujet en parallèle, par le biais de réseaux entièrement connectés entraînés sur le jeu de données MuCo-3DHP incorporant des données avec des occultations sévères. Enfin, l'ajustement cinématique du squelette est effectué en minimisant un terme d'énergie composé de la position (par cinématique inverse pour les caractéristiques 3D et la reprojection 2D), de l'orientation et de la cohérence temporelle. Il s'agit d'un système complexe qui implique de nombreuses étapes, ainsi qu'un re-traçage de chaque sujet avant la dernière étape. Cependant, le système est robuste aux occlusions et adapté à de multiples personnes. Il est également efficace sur le plan computationnel puisqu'il peut fonctionner en temps réel à 30 images par seconde sur des configurations matérielles génériques. Cependant, il semble que la dernière étape du système diminue l'exactitude globale tout en étant plus performante sur certaines articulations et en produisant une estimation de l'orientation plus lisse.

Zhou et al. (2017) ont publié une méthode en deux étapes utilisant l'architecture de réseau en sablier de Newell et al. (2016) pour la génération de cartes de chaleur 2D, puis une régression de la profondeur pour chaque articulation. En outre, ils appliquent un processus faiblement supervisé pour exploiter des images qui n'ont été étiquetées qu'avec des vérités terrain 2D. La prédiction de la profondeur est réalisée à partir de caractéristiques à différentes résolutions, qui sont extraites de plusieurs niveaux du réseau du sablier. L'entraînement faiblement supervisé est appliqué en utilisant des données étiquetées en 3D et en 2D. Une fonction de coût euclidienne est appliquée aux prédictions sur des images avec des vérités terrain 3D, tandis qu'une fonction de coût géométrique est utilisée lorsque seul un étiquetage 2D est disponible. Cette fonction de coût ajoute des contraintes à partir des rapports de longueur moyens des membres parmi des groupements d'os prédéfinis. La principale contribution de ce travail est la technique faiblement supervisée : les auteurs ont évalué si la contribution des données 2D améliorait les résultats sur la partie 2D du système ou sur l'ensemble de la tâche d'estimation 3D. Ils indiquent qu'à PCKh@0.5 (métrique standard pour l'estimation de la posture en 2D, voir 3.2.1), les résultats sont similaires pour l'estimation de la posture humaine en 2D lorsque les données 2D ne sont pas utilisées, mais l'estimation de la profondeur est grandement améliorée. Cependant, une analyse avec un seuil inférieur à 0,5 pourrait être intéressante, car une légère augmentation de l'exactitude 2D est tout de même observée. Néanmoins, leur travail confirme que l'utilisation de détecteurs 2D dans le cadre de la tâche de détection 3D est possible et donne des résultats précis ($\simeq$65mm MPJPE).

Martinez et al. (2017) présentent une méthode simple de référence pour l'estimation de la posture humaine en 3D. Cette méthode diffère des autres en ce qu'elle n'utilise pas de données d'image ou de cartes de caractéristiques intermédiaires (par exemple des cartes de chaleur de localisation des articulations), ni d'étapes d'optimisation avec ajustement du modèle. Au lieu de cela, elle déduit les coordonnées 3D des coordonnées 2D obtenues à l'aide d'une architecture d'estimation de posture humaine 2D de l'état de l'art. Malgré cette conception simple, il produit des résultats précis, comparables et parfois même meilleurs que certaines techniques contemporaines plus complexes. Les choix de conception du modèle sont les suivants : un simple réseau de convolution à deux couches avec normalisation par batch, activation ReLU et dropout. Il prend en entrée les prédictions 2D de Hourglass Network (Newell et al., 2016). Les résultats obtenus sur Human3.6M atteignent 62mm d'erreur moyenne sur des images uniques. Cette méthode est également l'une des premières à proposer l'exploitation directe des coordonnées 2D de détecteurs 2D efficients vers la 3D. La grande exactitude obtenue avec une approche simple, sans données d'image en entrée, a conduit les auteurs à émettre l'hypothèse que les caractéristiques visuelles utilisées dans les méthodes contemporaines n'étaient pas si utiles à l'estimation de la posture humaine en 3D ou étaient encore sous-exploitées. Cette dernière hypothèse tend à être confirmée par de nouvelles méthodes atteignant une exactitude croissante grâce à une exploitation intelligente des caractéristiques temporelles (Hossain and Little (2018), Pavllo et al. (2019), Cheng et al. (2019) et Cheng et al. (2020)). Parce qu'elle est simple, rapide et qu'elle utilise les performances croissantes des détecteurs 2D, cette méthode et sa technique en deux étapes ont inspiré de nombreuses études récentes.

Sun et al. (2018) a introduit la "régression intégrale" pour extraire les coordonnées 3D de cartes de confiance 2D. Il s'agit d'une fonction similaire à la fonction soft-argmax couramment utilisée en classification pour normaliser les sorties. Ici, elle est utilisée pour passer des cartes de pixels 2D à des coordonnées différentiables, permettant une régression directe dans un réseau. L'article décrit des études par ablation approfondies sur différentes procédures d'entraînement, fonctions de coût et architectures de base (réseaux en sablier et résiduels ; 3.3.3) et présente de bons résultats pour l'estimation de la posture humaine en 2D et 3D. Les auteurs adoptent également une stratégie d'apprentissage similaire à Sun et al. (2017) permettant l'utilisation de données étiquetées en 2D avec une supervision séparée pour les coordonnées xy et pour la profondeur, ce qui permet d'obtenir une exactitude encore plus élevée sur Human3.6M. De nombreuses approches utilisent désormais des cartes de chaleur en 2D avec la régression soft-argmax.

Omran et al. (2018) présentent l'approche "Neural Body Fitting" qui combine la segmentation de membres avec un réseau de neurones convolutif et un modèle paramétrique du corps humain (SMPL, Loper et al. (2015)). La première étape prédit une carte de segmentation des membres en utilisant l'architecture de réseau de neurones RefineNet (Lin et al., 2016). Dans la deuxième étape, les masques des membres produits sont directement transmis à un réseau ResNet-50 qui estime les paramètres de posture et de forme du modèle SMPL. L'ensemble du processus est entièrement différentiable et peut également être entraîné avec des données 2D de manière faiblement supervisée. Les auteurs le démontrent en re-projetant les coordonnées de l'articulation sur des images 2D et constatent qu'avec seulement 20% des données d'entraînement possédant des vérités terrain 3D, on obtient la même exactitude qu'avec des annotations complètes. Cette méthode est l'une des premières à décrire l'apprentissage d'un réseau de neurone suivi de l'adaptation à un modèle de corps humain paramétrique dans un système intégré unique. L'autre caractéristique qui différencie cette méthode des autres est son utilisation de carte de segmentation de membres en tant que caractéristiques intermédiaires. Selon l'analyse de l'auteur, l'utilisation d'une segmentation en 24 membres en entrée a conduit à des résultats significativement meilleurs qu'avec l'image directement ou les coordonnées des articulations. Ce résultat est particulièrement intéressant et devrait être pris en compte lors de la conception d'architectures qui veulent raisonner sur la structure 3D du corps humain.

De même, Kolotouros et al. (2019) a conçu un système avec l'utilisation conjointe d'un réseau de neurone convolutif pour la régression directe des points d'articulation et d'une technique d'optimisation itérative sur le même modèle volumétrique humain (SMPL, Loper et al. (2015)). Le réseau de neurone produit une bonne initialisation pour l'ajustement itératif du modèle humain. Ensuite, une fois que la position du modèle est affinée, elle est utilisée pour calculer la fonction de coût du réseau par rapport à la prédiction initiale, ce qui augmente son exactitude. L'originalité de la méthode consiste à utiliser le meilleur des deux techniques. La régression directe permet une initialisation rapide sans connaissance a priori directement sur les données de l'image ; l'optimisation itérative à partir d'un modèle humain produit une meilleure forme pour l'ajustement à l'image. Les deux parties du système se complètent et s'améliorent mutuellement au cours de chaque cycle d'apprentissage. Les résultats obtenus en utilisant cette méthode sur Human3.6M sont les plus précis des méthodes monoculaires non temporelles (sur une seule image RVB isolée). L'erreur moyenne est de 41,1 mm, ce qui est proche de l'exactitude des méthodes temporelles. Ces résultats montrent qu'avec des données d'entrée simples et une méthode adaptée, il est possible d'obtenir une grande exactitude. La question qui se pose est la suivante : cette méthode de "régression directe/adaptation de modèle" utilisant uniquement des caractéristiques visuelles pourrait-elle être complétée par des données temporelles afin d'atteindre un score d'exactitude plus élevé ?

Wandt and Rosenhahn (2019) utilise un espace de chaîne cinématique ("kinematic chain space") pour représenter la posture humaine en 3D dans leur réseau de discrimination (ou "critique"). Pour projeter les coordonnées de la posture dans l'espace de chaîne cinématique, les membres (c'est-à-dire les arêtes entre les point d'articulations détectées) sont décrits comme des vecteurs directionnels. Il est ensuite possible de retranscrire cette représentation en coordonnées de points. La représentation en chaîne cinématique contient des informations sur la longueur des membres, les angles et la symétrie du corps, tout en étant facile à calculer. Par la suite, Cheng et al. (2020) a utilisé une version temporelle de l'espace de la chaîne cinématique qui rend compte de la modification de l'angle et de la longueur à travers les images d'une vidéo.

Xu et al. (2019) ont proposé le système DenseRaC qui convertit des cartes "pixel-à-surface" du corps humain (IUV) en paramètres pour un modèle statistique du corps humain. Comme Omran et al. (2018), ce système utilise des caractéristiques intermédiaires mais au lieu de la segmentation des membres, DenseRaC utilise des cartes de surface "pixel-à-3D". Dans la première étape du pipeline, ces cartes sont calculées à l'aide de l'architecture Densepose-RCNN (un réseau entraîné sur Densepose-COCO, un jeu de données de surfaces corporelles humaines annotées manuellement). Pour améliorer l'entraînement, les auteurs présentent un jeu de données à grande échelle de postures humaines synthétiques qui peuvent également produire facilement des IUV. Dans un deuxième temps, les IUVs sont soumis à un réseau de régression qui estime les paramètres de forme et de posture du corps humain (similaire à Loper et al. (2015)) ainsi que les paramètres de la caméra. Une fois reconstruit, le maillage 3D est reprojeté à l'aide d'un moteur de rendu différentiable et rastérisé dans un IUV similaire à ceux produits dans la première étape. Ensuite, le calcul d'une fonction de coût antagoniste avec un réseau discriminant permet d'éliminer les configurations impossibles.

Kocabas et al. (2019b) tire parti de la configuration multivues que fournit chaque jeu de données de capture de mouvement. Les auteurs proposent d'inférer la géométrie à partir de la détection 2D correspondante dans chaque vue, puis de déduire les coordonnées 3D. Cette technique rend l'auto-supervision possible avec des données non étiquetées. Contrairement à de nombreuses techniques multivues (Iskakov et al. (2019), He et al. (2020)), cette méthode n'utilise pas les paramètres de la caméra. Il faut noter que lors de l'entraînement, le modèle utilise les images multivues, mais pendant la prédiction, il devient une méthode monoculaire. L'architecture est composée de deux réseaux utilisant ResNet50 et une couche de déconvolution comme architecture de base (voir architectures 2D : 3.3.3). Les deux réseaux produisent des cartes de chaleur spatiales pour chaque articulation dans chaque vue. La différence est que, dans un cas, elles sont converties en coordonnées 2D et en 3D dans l'autre (toutes

deux avec une fonction soft-argmax). Le réseau 3D est celui qui sera utilisé pour les inférences, et qui a des poids qui peuvent être appris. Les poids du réseau 2D sont gelés et seront utilisés pour l'auto-supervision. Bien que les caméras ne soient pas calibrées, les auteurs conseillent de détecter les points clés des articulations dans chaque vue pour obtenir les paramètres de la caméra (en utilisant RANSAC et SVD). Ensuite, la triangulation peut être effectuée pour obtenir les coordonnées 3D, qui sont ensuite utilisées pour superviser le réseau 3D avec une fonction de coût absolue lissée. Le score d'exactitude de la méthode est actuellement le meilleur lorsque peu ou pas de données étiquetées sont disponibles (70 mm de moyenne MPJPE sur Human3.6M et 64,7 3DPCK sur MPI-INF-3DHP). Ce schéma d'apprentissage est prometteur et permet d'obtenir une grande exactitude en ne nécessitant que des données brutes (sans tenir compte du détecteur 2D pré-entraîné), à condition que des images multivues soient accessibles.

Enfin, il est important de considérer les méthodes utilisées dans plusieurs domaines de recherche. La principale, DeepLabCut (Mathis et al. (2018), Mathis and Mathis (2019), Nath et al. (2018)) est un système générique de suivi de points clés sans marqueur développé dans le but d'obtenir des résultats similaires à la précision de l'annotation humaine. Sa cible d'application initiale était le suivi vidéo de points clés prédéfinis pour différentes espèces animales. DeepLabCut obtient de bons résultats avec peu de données d'entraînement, ce qui l'a rendu populaire et il est maintenant cité dans de nombreux articles en neurosciences et en recherche sur le mouvement humain. Pour l'analyse de la marche humaine (Moro et al. (2020), Fiker et al. (2020)), son exactitude est similaire à celle de la capture de mouvement basée sur des marqueurs. Bien que DeepLabCut ne soit pas au sens strict un modèle d'estimation de posture humaine, il peut être utilisé comme tel. En effet, il est basé sur une architecture d'estimation de posture humaine : Il est constitué des premières couches du réseau DeeperCut (Insafutdinov et al., 2016), qui est un modèle d'estimation de posture 2D multipersonnes utilisé ici pour l'extraction de caractéristiques, suivi d'une couche de déconvolution. Ce modèle et son inspiration ne relèvent pas directement de l'estimation de la posture humaine en 3D. Pourtant, DeepLabCut a été utilisé dans un contexte de caméras calibrées multivues avec une simple triangulation (Nath et al. (2018), Sheshadri et al. (2020)). Comme cette méthode semble populaire dans différents domaines pour sa flexibilité et ses performances dans une grande variété de contextes, il est important de noter que des résultats plus récents sur la détection de points clés 2D spécifiques ont surpassé DeeperCut. Le principal attrait de cette méthode est la définition d'étiquettes personnalisées et sa puissante capacité de généralisation.

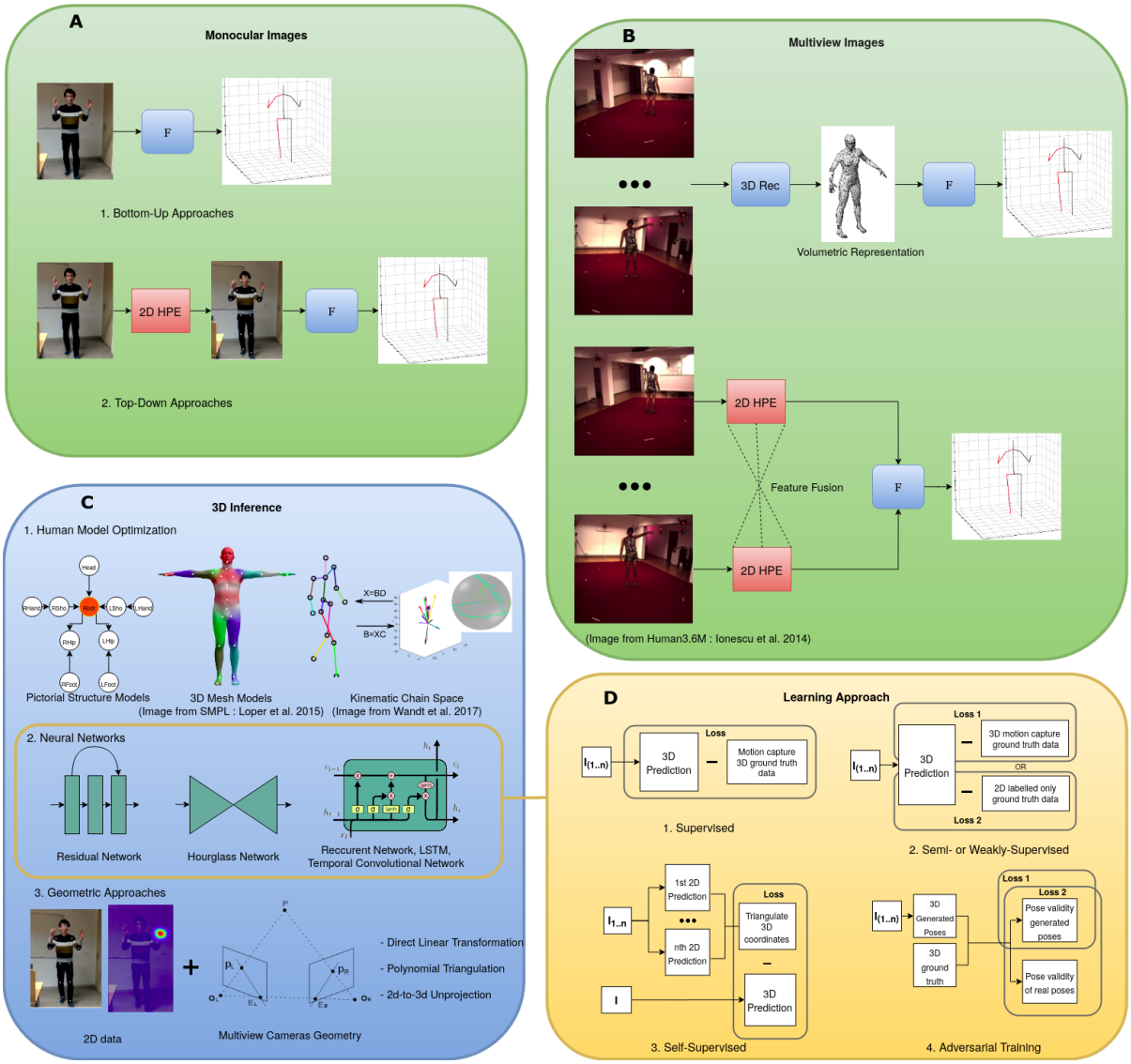


FIGURE 3.2 : Aperçu des différents niveaux d'estimation de la posture humaine en 3D sans marqueur. **A** : Approches monoculaires, les architectures de base d'estimation de posture 2D couramment utilisées sont décrites dans 3.3.3. **B** : Exploitation des caractéristiques 3D et détection 2D multivues comme entrée pour les détecteurs 3D. **C** : Les différentes familles d'estimation de posture 3D. **D** : exemples d'approches d'apprentissage appliquées à l'estimation de la posture humaine.

3.4.2 Séquence d'images monoculaires

Méthode	Human3.6M	MPI-INF-3DHP	HumanEva
Hossain and Little (2018)	58.5	-	22.0
Cai et al. (2019)	48.8 / 39.0*	-	-
Pavllo et al. (2019)	46.8 / 36.5*	-	23.1/15.8*
Cheng et al. (2019)	42.9/32.8*	-	14.3*
Cheng et al. (2020)	**40.1**	**84.1**	**13.5***
Liu et al. (2020)	45.1 / 35.6*	-	15.4*
Wang et al. (2020)	42.6 / 32.7*	86.9(2d GT)	-

TABLE 3.4 : Comparaison de l'exactitude de plusieurs méthodes temporelles monoculaires de l'état de l'art. Les résultats de Human3.6M et HumanEva sont rapportés en MPJPE absolue (la plus basse est la meilleure) ; les résultats de MPI-INF-3DHP sont rapportés en 3DPCK (la plus haute est la meilleure). Les techniques avec l'annotation + utilisent des données d'entraînement supplémentaires pour obtenir le résultat ; les autres utilisent les protocoles recommandés par le jeu de données de référence. L'annotation * indique les résultats publiés avec alignement par procrustes sur les postures de vérité terrain avant évaluation.

Hossain and Little (2018) ont utilisé l'architecture séquence-à-séquence qui a été initialement appliquée dans les tâches de traduction automatique pour l'estimation de la posture humaine. L'idée principale est d'utiliser le contexte temporel à long terme pour prédire une nouvelle séquence, comme pour la traduction d'un texte dans une autre langue. Ici, les séquences de coordonnées 2D (prédites avec l'architecture de réseau en sablier de Newell et al. (2016)) sont utilisées pour prédire les séquences 3D en utilisant des "Long Short-Term Memory units" (LSTM). La première couche du réseau encode les séquences dans un espace caché et la dernière couche les décode en une séquence tridimensionnelle en utilisant des connexions résiduelles. De plus, des choix de conception importants sont faits pour l'entraînement. Premièrement, la dérivée première des coordonnées 3D est incluse dans la fonction de coût afin d'atténuer l'effet des erreurs durant l'étape de prédiction 2D en renforçant la cohérence temporelle. Ensuite, l'entraînement combine la normalisation des couches, du "dropout" récurrent et enfin la pondération des ensembles d'articulations, afin de forcer le réseau à mieux prédire des membres difficilement détectable. À ce jour, la méthode est l'une des plus précises parmi les méthodes utilisant des données temporelles. Notamment, une évaluation réalisée par les auteurs montre que la longueur de séquence optimale pour leur modèle est de 5 images. Au-delà de ce point, l'exactitude diminue lentement, ce qui suggère que le contexte temporel à long terme ne fournit pas de bonnes informations ou que le modèle n'exploite pas correctement les données d'image passées.Un

problème de cette architecture est la longueur fixe des données temporelles d'entrée, qui a été résolu plus tard en utilisant des réseaux convolutifs temporels (TCN) (Pavllo et al. (2019), Cheng et al. (2020). Cet article détaille et analyse d'autres techniques utiles appliquées aux données temporelles de posture, telles que la contribution importante des connexions résiduelles.

Cai et al. (2019) utilise également la dépendance temporelle dans la séquence de posture qui compose le mouvement du sujet. Cependant, en plus des contraintes temporelles, les contraintes de position des articulations sont également appliquées à l'aide d'une structure de graphe. Des arêtes sont présentes pour modéliser la continuité spatiale et la symétrie, mais aussi les connexions avec la même articulation dans les images passées et futures de la vidéo. Ces graphes sont calculés à partir de squelettes 2D qui sont ensuite transmis à un réseau convolutif de graphes (GCNN). En fonction du type de voisinage (diverses contraintes spatiales et connexions temporelles), les auteurs proposent un type de convolution différent. Ensuite, une architecture locale-à-globale est utilisée de manière similaire à Newell et al. (2016) pour obtenir des prédictions 3D. La structure du graphe est successivement sous- puis sur-échantillonnée à l'aide de caractéristiques de niveau inférieur provenant des couches précédentes du réseau de manière hiérarchique. Ensuite, la méthode produit une posture 3D à partir d'un module de raffinement de la posture constitué d'une couche entièrement connectée. Les auteurs utilisent une stratégie de coûts combinant différents termes pour renforcer la localisation des articulations, la symétrie des membres et la fluidité temporelle. Les résultats obtenus sur différents jeux de données de référence montrent une meilleure exactitude que les méthodes basées uniquement sur des séquences temporelles. Cependant, la méthode n'est probablement pas adaptée aux séquences inférieures à 7 images pour la plupart des actions. Il est également important de noter que l'effet du détecteur 2D choisi n'a pas été testé.

Comme mentionné plus haut, Pavllo et al. (2019) est une technique qui consiste en une méthode monoculaire qui exploite un réseau entièrement convolutif pour traiter les données temporelles. L'avantage principal étant que leur architecture ne nécessite pas de donnée en entrée de taille fixe contrairement aux réseaux de neurone récurrent (RNN) et au LSTM. Dans cet article, les auteurs présentent une architecture basée sur des convolutions dilatées dans le temps qui applique également la semi-supervision pour ajouter des données non étiquetées à l'entraînement. Les convolutions dilatées capturent les dépendances à long terme, mais ont également besoin de moins de paramètres d'apprentissage et ont une meilleure vitesse de calcul que les modèles séquence-à-séquence (Hossain and Little, 2018). Les auteurs présentent une comparaison montrant que leur architecture nécessite environ le même nombre de paramètres et de FLOPs (floating-point operations per second) tout en obtenant une meilleure exactitude. Le processus semi-supervisé qu'ils ont utilisé est appelé "back-

projection" : une architecture d'encodeur-décodeur encode une prédiction d'articulation 2D en posture 3D et la décode en la projetant à nouveau en 2D dans l'image. L'erreur est ensuite calculée entre la prédiction originale et la prédiction 2D "rétroprojetée". Cette méthode permet d'augmenter l'exactitude de 10% sur deux des principaux jeux de données de référence, ce qui conduit à une erreur moyenne inférieure à 50 mm pour une méthode monoculaire. Les méthodes qui suivent cet article, basées sur des séquences vidéo, utilisent toujours l'architecture TCN (Temporelle Convolutional Networks) (Cheng et al. (2019), Cheng et al. (2020), Liu et al. (2020)) introduite ici. Un autre résultat intéressant est l'impact du détecteur 2D utilisé. Les tests effectués montrent que les détecteurs 2D CPN et Mask-RCNN donnent de meilleurs résultats en définitive avec leur modèle. Les auteurs pensent que "cela est dû à la résolution plus élevée des cartes de chaleur et à la combinaison de caractéristiques plus fortes" de ces détecteurs.

Cheng et al. (2019) et Cheng et al. (2020) utilisent également des réseaux de neurone convolutifs temporels avec un entraînement spécifique à l'occultation pour améliorer la prédiction sur des images difficiles.

Dans Cheng et al. (2019), ils décrivent un réseau sensible à l'occultation où un "modèle d'homme-cylindre" produit des étiquettes d'occultation. Leur architecture est composée d'un détecteur de personne entraînable, d'un estimateur de posture 2D, d'un estimateur de posture 3D et enfin d'un discriminateur de posture. Les deux algorithmes d'estimation de la posture sont respectivement un réseau convolutif temporel 2D et 3D. Ils utilisent également des données uniquement en 2D pendant l'entraînement en utilisant une fonction de coût de reprojection comme dans Pavllo et al. (2019). Le modèle d'occultation intervient à ce stade quand les points auto-occultés (calculés avec le modèle cylindrique) ne sont pas pris en compte, car ils sont jugés non fiables.

Sur la base des travaux précédents, Cheng et al. (2020) décrit un modèle entraînable de bout en bout avec plusieurs modules qui reconstruisent les articulations 3D à partir de vidéos monoculaires. Des cartes de chaleur de confiance en 2D sont estimées et utilisées comme caractéristique pour la prédiction 3D. Ils utilisent un réseau convolutif multiéchelle (HRNet) qui fusionne les caractéristiques spatiales. Ainsi, les principales caractéristiques de leur méthode sont les suivantes :

- Apprentissage d'un plongement multiéchelle obtenu à partir de ces cartes de chaleur. Les postures 3D sont ensuite prédites à l'aide de ce plongement en utilisant un réseau convolutif temporel (TCN) (Pavllo et al., 2019).
- Validation des séquences de posture à l'aide d'un modèle discriminant basé sur les chaînes cinématiques spatio-temporelles (qui impose des contraintes d'angle et de longueur aux membres).

- Augmentation des données en utilisant une occultation synthétique à différents niveaux pendant l'entraînement du TCN.

L'apprentissage semi-supervisé est utilisé dans le but d'inclure des données étiquetées strictement 2D pendant le processus d'entraînement comme dans Pavllo et al. (2019). Enfin, et uniquement pour l'étape d'entraînement, les vues multiples de la base de données Human3.6M sont utilisées pour assurer une bonne prédiction de l'orientation du squelette. Pour ce faire, les auteurs utilisent une fonction de coût comparant chaque inférence de deux paires de vues différentes au même moment dans la vidéo (après avoir appliqué une rotation de la caméra connue à partir des données de calibration). Parmi les méthodes examinées ici, cette méthode obtient les meilleurs résultats globaux sur HumanEVA et MPI-INF-3DHP. Elle obtient également les meilleurs résultats pour une méthode monoculaire sur Human3.6M (MPJPE 40mm). Cependant, il s'agit d'une méthode complexe avec de nombreux sous-modules et paramètres réduisant les erreurs dues à l'occultation. Elle exploite et étend également le discriminateur de posture basé sur l'espace de chaîne cinématique de Wandt and Rosenhahn (2019), en l'utilisant avec des données temporelles. Les auteurs abordent la contribution de chaque module à l'exactitude finale. Cependant, ils le font en ajoutant les modules un par un à l'architecture principale, ce qui ne permet pas de savoir si un module améliore ou compense les erreurs ou les performances d'un autre. Une comparaison croisée plus poussée pourrait aider à déterminer quel module a le plus d'impact et sur quelles données ou dans quel contexte.

Liu et al. (2020) Ajoutent un mécanisme d'attention à l'approche temporelle pour l'extraction de la posture 3D à partir de la vidéo. Comme pour (Pavllo et al., 2019), la convolution dilatée temporelle extrait des informations dans la séquence de postures 2D. Le mécanisme d'attention ajouté sélectionne les images et tenseurs de sortie qui sont les plus utiles pour la détection. Les modules d'attention temporelle sont calculés à partir de la distribution des tenseurs à chaque étape de l'entraînement. Les modules d'attention du noyau de convolution sont calculés à partir de la distribution des sorties des canaux de chaque couche. Les deux modules d'attention sont propagés dans une matrice d'attention à la couche suivante. Par-dessus cette architecture, des convolutions dilatées multiéchelles (voir la convolution de dilaté dans la méthode décrite plus haut Pavllo et al. (2019)) avec des champs réceptifs croissants sont utilisées pour réduire les problèmes de fuite de gradient. Cette méthode montre des progrès par rapport aux méthodes les plus récentes. Une étude d'ablation montre également comment les modules d'attention associés à la stratégie de convolution multiéchelle permettent d'obtenir de meilleurs résultats, notamment sur les images difficiles (mouvements rapides ou sujets partiellement occultés).

Les deux principales contributions de Wang et al. (2020) sont : "*Motion loss*", une nouvelle fonction de coût basée sur la trajectoire des points clé et "*U-shaped Graph*

Convolutional Network" (UGCN), une nouvelle architecture de réseau de neurone convolutif de graphe. La fonction de coût de mouvement est basée sur les vecteurs de coordonnées. Pour la rendre différentiable, afin qu'elle puisse être utilisée dans une architecture d'apprentissage, toute séquence de mouvement doit être codée à l'aide d'un opérateur différentiable. Les auteurs ont choisi empiriquement le produit scalaire (parmi d'autres opérateurs testés). La fonction de coût finale est composée de cette "Motion Loss" calculée avec le "codage de mouvement par paire" et d'une fonction de coût de reconstruction de position absolue. L'UGCN utilise un graphe spatio-temporel pour représenter les mouvements. Ensuite, des convolutions spatiales sont appliquées sur chaque squelette pour chaque image avant que des convolutions temporelles soient appliquées à la dimension temporelle de chaque articulation dans le graphe. L'architecture du réseau est similaire à celles qui ont fait leurs preuves en segmentation sémantique (par exemple : Ronneberger et al. (2015). Elle consiste en trois étapes : le sous-échantillonnage, le sur-échantillonnage et la fusion. Cette architecture capture les caractéristiques à différentes échelles, ce qui implique que le UGCN explore les informations temporelles et spatiales à différentes échelles. Cette architecture a été testée en ajoutant les sur-échantillonnages et les sous-échantillonnages un par un, montrant une exactitude croissante. L'ajout de la "fonction de coût de mouvement" améliore aussi considérablement les résultats sur deux jeux de données de référence (Human3.6M et MPI-INF3DHP), améliorant ainsi les résultats de l'état de l'art.

3.4.3 Multivue

Qiu et al. (2019) utilise un processus de prédiction en deux étapes similaire à Zhang et al. (2020) sans prendre des données d'IMU comme entrées. Au lieu de cela, ils choisissent de fusionner les données des images multivues en utilisant des contraintes de géométrie projective. Ce processus est réalisé par une couche convolutive qui fusionne les données de pixels de chaque vue le long des lignes épipolaires en utilisant des matrices de poids. Après cette étape, des cartes de chaleur des positions des articulations 2D fusionnées sont générées et la posture 3D est inférée par le modèle de "Pictorial Structure" avec un modèle de squelette d'articulation. Une variation récursive de ce modèle est utilisée pour réduire les erreurs de quantification et la complexité en utilisant un algorithme diviser pour régner pour la discrétisation de l'espace. En comparant les résultats des méthodes fournissant des coordonnées absolues, cette approche de vues croisées a amélioré les résultats de l'état de l'art à l'époque et les améliore d'autant plus en utilisant des données 2D supplémentaires pendant l'entraînement. Elle donne également des résultats compétitifs sur le jeu de données TotalCapture sans entraînement préalable et sans utiliser les données IMU. Elle peut également être utilisée dans différentes configurations de caméras en utilisant un pseudo-étiquetage à partir de l'estimateur de posture 2D. La méthode peut être appliquée à l'estimation de la posture humaine en 3D dans de nouveaux contextes sans avoir besoin d'entraînement

Méthode	Human3.6M	Total Capture	Input
Qiu et al. (2019)	31.17 / 26.21+	29	Multivues
Iskakov et al. (2019)	**17.7+**/20.80*+	-	Multivues
He et al. (2020)	26.9/19.0+	-	Multivues
von Marcard et al. (2016)	-	-	Multivues, IMU
Trumble et al. (2017)	87.3	77.0	Multivues, Temporelle, IMU
von Marcard et al. (2018)	-	26.0	Monoculaire, IMU
Huang et al. (2019)	37.5/13.4*	28.9	Multivues, IMU
Zhang et al. (2020)	-	**24.6**	Multivues, IMU

TABLE 3.5 : Comparaison de l'exactitude de plusieurs méthodes multivues et multimodales de l'état de l'art. Les résultats de Human3.6M et TotalCapture sont rapportés en MPJPE absolue (la plus basse est la meilleure). Les techniques avec l'annotation + utilisent des données d'entraînement supplémentaires pour obtenir le résultat ; les autres utilisent les protocoles recommandés par le jeu de données de référence. L'annotation * indique les résultats publiés avec alignement par procrustes sur les postures de vérité terrain avant évaluation.

avec des vérités terrain en 3D. Cette méthode illustre très bien l'approche principale qui consiste à utiliser les résultats impressionnants en estimation de posture humaine 2D et à les transposer en 3D. Ici, l'affinement des prédictions est effectué en ajoutant des vues multiples pour améliorer les prédictions 2D.

Iskakov et al. (2019) présente une méthode entraînable pour trianguler les postures humaines. Cet article propose deux méthodes géométriques pour trianguler les coordonnées des articulations 3D à partir de cartes de chaleur des articulations dans des vues multiples. La première est une méthode algébrique basée sur la résolution d'un système d'équations vectorielles avec des coordonnées 3D. La seconde méthode est une triangulation à partir de l'agrégation volumétrique des reprojections des cartes de chaleur 2D dans une grille de voxels. Les deux méthodes pondèrent les informations provenant de différentes vues avec des coefficients appris. Dans l'approche volumétrique, chaque carte de chaleur correspondant à une articulation dans chaque vue différente est projetée dans un cube de voxels. Ces cartes volumétriques (pour une articulation spécifique) sont ensuite agrégées avec une pondération de l'impact des différentes vues puis passée dans un réseau de convolution 3D qui les affine. L'étape finale consiste en une opération soft-argmax sur les cartes de chaleur 3D résultantes, ce qui donne des coordonnées 3D calculables. Chacune de ces étapes est différentiable et les poids des différentes couches convolutionnelles à chaque étape sont mis à jour en utilisant une fonction de coût absolue. Les résultats sur Human3.6M pour une entrée à vues multiples sont les meilleurs à ce jour avec une MPJPE de 17,7 mm (c'est-à-dire

avec la méthode volumétrique et l'agrégation softmax pendant l'étape d'apprentissage comme décrit ci-dessus). Une des limites de cette méthode est qu'elle a besoin d'un volume correctement cadré autour du squelette humain pour bien fonctionner. Par conséquent, au moins deux caméras doivent détecter l'articulation du bassin. Le score absolu en MPJPE donné est également calculé en supprimant plusieurs actions dues à des erreurs d'annotation sur Human3.6M. Néanmoins, cette méthode atteint la meilleure exactitude pour l'estimation de la posture relative parmi toutes les méthodes examinées.

Les "Epipolar Transformers" de He et al. (2020) sont des modules utilisant le mécanisme d'attention ainsi que la connaissance de la géométrie épipolaire. Les auteurs ont remarqué que la plupart des détections de points clés en 2D n'utilisaient pas du tout les caractéristiques 3D : c'est ce qui a motivé leur "Epipolar Transformer". L'objectif est de fusionner les caractéristiques intermédiaires de plusieurs vues pendant l'inférence 2D en utilisant les contraintes de la géométrie projective comme Qiu et al. (2019). Tout d'abord, à partir d'un point détecté dans la vue source, le module échantillonne tous les points sur la ligne épipolaire correspondante dans une autre vue. Ensuite, les caractéristiques de cette ligne sont fusionnées en fonction d'une similarité pondérée calculée avec le point source. Au final, les cartes de caractéristiques obtenues ont la même taille que l'entrée, ce qui rend le module compatible avec n'importe quel système multivues à deux niveaux. Tout algorithme de triangulation peut alors être appliqué. Sur Human3.6M, les auteurs ont calculé les coordonnées 3D avec le modèle de "pictorial structure" récursif de Qiu et al. (2019) en utilisant leur module : ils ont obtenu le meilleur score MPJPE de l'état de l'art, et ce, sans utiliser de données d'entraînement externes. Ils comparent également leurs résultats avec un entraînement préalable sur le jeu de données MPII 2D et obtiennent des résultats proches des meilleures méthodes (19 mm contre 17,7 pour Iskakov et al. (2019)) avec 10% de paramètres et d'opérations de calcul en moins. La principale limite de la méthode est qu'elle ne fonctionne que sur un système multicaméras entièrement calibré, car elle a besoin des paramètres des caméras pour calculer les lignes épipolaires.

3.4.4 Approches Multimodales

von Marcard et al. (2016) proposent l'une des premières méthodes combinant la vidéo multivues et des IMU. Les auteurs affirment que leur méthode est moins intrusive que la MoCap basée sur les marqueurs, car seuls quelques capteurs sont placés sur les sujets. Ils utilisent une représentation des contraintes du corps humain basée sur des chaînes cinématiques. Avec les silhouettes des vues multiples extraites par soustraction de l'arrière-plan et l'orientation des membres à partir de l'IMU, ils minimisent un terme d'énergie hybride obtenu à partir de l'orientation et de la cohérence des contours avec un modèle de maillage humain (Loper et al. (2015)). Ils fournissent une

analyse approfondie de leur méthode (avec le jeu de données TNT15 présenté dans le même article ainsi que sur HumanEva Sigal et al. (2010)) montrant que la vidéo et le capteur d'orientation se complètent. L'idée est que les IMU mesurent avec exactitude les angles des articulations, mais ont tendance à dériver au cours de l'expérience, alors que la vidéo est mieux adaptée pour obtenir des informations de position des articulations.

Trumble et al. (2017) présentent le premier jeu de données de capture de mouvement à grande échelle qui contient également des données IMU (TotalCapture). L'article décrit une technique d'estimation de la posture en 3D qui fusionne des données 3D provenant de vues multiples et l'orientation des membres provenant de l'IMU, tout en maintenant le contexte temporel à l'aide d'une couche LSTM sur les cinq précédentes images de la vidéo. Une "Probabilistic Visual Hull" (Grauman et al., 2003) est calculée à partir de la vidéo multivues et introduite dans un réseau convolutif 3D avec 26 coordonnées d'articulation 3D en sortie. En parallèle, un processus la résolution cinématique fournit les mêmes coordonnées articulaires à partir des données IMU. Les vecteurs des deux sources sont ensuite transmis à la couche LSTM et fusionnés en un plongement représentant la posture 3D. De cette manière, les auteurs montrent qu'il est possible d'apprendre une "correspondance entre les estimations d'articulation prédites par les deux sources de données et les emplacements réels des articulations". Ils évaluent leur méthode sur le nouvel jeu de données TotalCapture et Human3.6M.

Suite à leur travail avec les unités inertielles, von Marcard et al. (2018) présentent une méthode monoculaire avec une caméra mobile dans la nature et plusieurs sujets portant des IMUs. La méthode utilise un algorithme d'estimation de posture multi-personnes 2D (Cao et al., 2017) et un module qui ajuste le modèle SMPL (Loper et al., 2015) aux données IMU. Ensuite, chaque squelette 2D est associé à une posture et une forme 3D à l'aide d'une optimisation de graphe. Ensuite, avec les paramètres du modèle d'association obtenus, la posture et l'orientation de la caméra sont optimisées et réinjectées pour les itérations suivantes. Évaluée sur l'ensemble des données de TotalCapture, la méthode surpasse les précédentes de 44 mm. Cette technique permet de capturer un nouveau jeu de données dans des environnements extérieurs et sans vérité terrain de marquage : le jeu de données "3D Poses in the Wild" décrit dans l'article.

Huang et al. (2019) ont développé un réseau convolutif 3D entraînable de bout en bout avec un module de raffinement basé sur les données IMU. L'idée principale est de traiter des données 3D à partir d'images multivues sans aucune transformation. Un volume multicanal, construit à partir des silhouettes humaines segmentées et des paramètres de la caméra, est utilisé comme entrée pour le réseau. Des cartes de chaleur 3D de confiance en voxels sont calculées à cette phase et peuvent déjà être utilisées pour prédire l'estimation de la posture humaine. L'étape de raffinement fusionne les

volumes construits à partir de l'IMU et ceux construits à partir des cartes de chaleur, ainsi que les volumes multicanaux. Les volumes IMU sont transformés en un cylindre représentant des membres à partir de l'orientation en quaternions et de la position des articulations prédite précédemment. Ensuite, toutes ces informations 3D sont introduites dans un autre réseau convolutif 3D. L'architecture utilise des réseaux en sablier (Newell et al., 2016) et des modules de réseaux résiduels (voir architectures 2D 3.3.3) et une fonction soft-argmax 3D pour extraire les coordonnées. Une fonction de coût EQM est calculée aux différents niveaux de l'architecture. En coupant aléatoirement les informations provenant d'une caméra, la méthode permet d'augmenter les données pendant l'entraînement, ce qui améliore considérablement les performances sur les images partiellement capturées. Avec seulement leur module basé sur les données visuelles sur un jeu de données sans IMU, la méthode obtient une bonne exactitude globale. Les auteurs affirment que leur modèle peut être utilisé dans un système en temps réel, car il n'utilise pas de séquences temporelles. Cependant, ils ne fournissent pas d'évaluation de sa performance ou de sa vitesse. Ils soulignent également que leur technique n'utilise pas un modèle humain complexe et est donc plus susceptible de se généraliser à de nouveaux sujets. Cependant, cela dépend de la performance de l'algorithme de segmentation de la forme humaine à l'étape du pré-traitement.

De même, Zhang et al. (2020) fusionne l'IMU avec des données d'images multivues, mais cette fois en utilisant un modèle de squelette humain. L'approche consiste à affiner les cartes de chaleur de confiance des articulations en 2D avec les données IMU pour ensuite utiliser un modèle des membres du corps. Tout d'abord, comme dans Qiu et al. (2019), ils extraient des cartes de chaleur articulaires 2D. Cependant, les auteurs utilisent l'architecture de base de Xiao et al. (2018) pour la partie 2D avant de fusionner les informations des cartes de chaleur et l'orientation des IMU pour estimer de façon géométrique les coordonnées correctes. Ils appellent ce processus Orientation Regularized Network (ORN). Cette technique peut être utilisée avec n'importe quelle méthode de prédiction basée sur les cartes de chaleur et ils l'utilisent pour entraîner un réseau de bout en bout qui produit des résultats plus précis que les méthodes 2D de pointe n'exploitant pas les données IMU. La deuxième partie de leur travail consiste en une variante du Pictorial Structure Model (PSM) (Felzenszwalb and Huttenlocher (2005), Belagiannis et al. (2014)) qui est souvent utilisé pour l'estimation de posture 3D et 2D. Cette famille de méthodes calcule la posture humaine la plus probable parmi toutes celles possibles dans un espace discret. Typiquement, les méthodes de type PSM utilisent la longueur des membres comme contrainte principale. Ici, les auteurs exploitent les données IMU pour appliquer également une contrainte d'orientation des membres, ce qui améliore l'exactitude. Ce système comprend un module qui permet d'améliorer considérablement les résultats des estimateurs 2D avec des données inertielles. Notez que la partie 3D de ce système n'utilise pas de réseau de neurone convolutif, contrairement à de nombreuses autres méthodes actuelles, bien

qu'elle donne une excellente exactitude. Les auteurs utilisent également des données IMU synthétiques pour prédire la posture 3D sur le jeu de données Human3.6M pour conclure que leur utilisation améliore les résultats de 10 mm en moyenne.

Enfin, des travaux récents utilisant un seul capteur de profondeur Yu et al. (2017) et Yu et al. (2018) ont montré de bons résultats pour la capture de mouvement en temps réel. La première utilise l'algorithme "*Iterative Closest Point*" (ICP) pour reconstruire les formes du corps et la seconde optimise le modèle de corps humain SMPL. Cependant, ces méthodes ne sont pas testées avec des métriques d'erreur de localisation des articulations sur un jeu de données de référence d'estimation de la posture.

3.5 Analyse des critères de performance

L'analyse de la performance des méthodes d'estimation de posture est à mettre en parallèle avec les spécifications et pré-requis d'un système ainsi qu'avec son contexte d'utilisation. Ceux-ci peuvent varier en fonction du domaine d'application. Aujourd'hui, la posture humaine en 3D est utilisée dans de nombreux domaines :

Interface Homme Machine : Il existe un nombre croissant d'applications utilisant la pose et le geste humain pour interagir avec les ordinateurs. L'estimation de posture 3D est essentielle pour aider les robots et les machines à mieux comprendre et répondre aux mouvements humains.

Sécurité : L'application classique consiste à suivre les personnes dans des environnements intérieurs et extérieurs pour s'assurer qu'elles ne commettent pas de vol ou d'infraction.

Analyse du mouvement : Il s'agit d'un vaste domaine qui comprend l'analyse du sport et des performances, les études médicales, la sémantique des poses ou l'étude des interactions inter-humaines.

Divertissement : L'estimation de la pose peut être utilisée pour le contrôle d'un avatar (par exemple, Kinect) ou pour affiner les environnements en réalité virtuelle et augmentée, ainsi que pour l'animation d'avatars dans les jeux vidéos et les films.

Moeslund and Granum (2001) les classent ces domaines en trois catégories : surveillance, analyse et contrôle. Chaque catégorie exige des performances différentes en matière d'exactitude, de rapidité ou de robustesse. Cependant, au sein d'une même catégorie, il existe des variations sur ces critères. Par exemple, comme l'indiquent ces auteurs, une application de contrôle peut être limitée à un environnement hautement

contrôlé (contrôle d'un avatar) ou à une scène extérieure générique avec des conditions variables.

Pour cette raison, chaque critère de performance sera décrit séparément et il sera expliqué dans quel cas d'utilisation il est le plus pertinent. Les tableaux 3.6, 3.7 et 3.10 classent les méthodes en fonction de l'exactitude rapportée dans MPJPE. Le niveau de robustesse correspond au nombre d'hypothèses ou de contraintes nécessaires pour une détection correcte. Enfin, pour exprimer la rapidité, nous indiquons si le modèle peut s'exécuter en temps réel. Chacun de ces tableaux permet de faire une comparaison croisée de ces critères pour définir les méthodes les mieux adaptées aux critères les plus nécessaires.

3.5.1 Exactitude

L'exactitude est le principal critère utilisé pour évaluer les méthodes d'estimation de la pose humaine sans marqueur dans la communauté de la vision par ordinateur. Cependant, il présente certaines limites qu'il est important de garder à l'esprit. Tout d'abord, la plupart des jeux de données de référence comparent les méthodes de vision sans marqueur aux résultats obtenus par des systèmes optoélectroniques à base de marqueurs qui ne sont pas eux-mêmes exempts d'erreurs. Comme indiqué dans la section 3.2.2, certains anciens jeux de données de capture de mouvement contiennent des vidéos à faible résolution et des annotations inexactes. De plus, Colyer et al. (2018) rapportent une erreur d'en moyenne 10mm par rapport à la position réelle des articulations.

La deuxième limite dans la comparaison de l'exactitude est que, dans de nombreux cas, ce n'est pas le taux d'erreur local qui est important pour l'application, mais la sémantique de la pose dans son ensemble (Voir 3.2.1). Cela peut être le cas pour les systèmes de surveillance ou les interfaces homme-machine par exemple. Lorsque l'exigence d'exactitude (du moins avec les métriques conventionnelles) est faible ou minimale, il peut être judicieux de se pencher sur les deux autres critères discutés.

Cependant, de nombreux autres domaines de recherche nécessitent une grande exactitude, comme la médecine ou la biologie. En général, la condition préalable d'exactitude peut également être différente selon l'analyse. Parfois, la plus grande exactitude possible est requise, quelle que soit la complexité de la procédure d'acquisition et de calcul (par exemple, pour la recherche en biomécanique). Cependant, de nombreuses études sont menées à partir de vidéos étiquetées par des humains avec des configurations beaucoup plus simples. Nous proposons une solution basée sur la vision par ordinateur pour chacun de ces scénarios.

Exactitude				
Méthode	Type	Exactitude	Niveau de robustesse	Temps réel
Kolotouros et al. (2019)	Monoculaire	**41.1**	moyen	✗
Wandt and Rosenhahn (2019)	Monoculaire	<u>50.9</u>	moyen	✓
Iskakov et al. (2019)	Multivues	**17.7**	moyen	✗
He et al. (2020)	Multivues	<u>26.9</u>	faible	✗
Cheng et al. (2020)	Temporelle	**40.1**	moyen	✗
Wang et al. (2020)	Temporelle	<u>42.6</u>	moyen	✗
Zhang et al. (2020)	Multimodal	*24.6**	faible	✗

TABLE 3.6 : Méthodes les plus exactes. Les méthodes sont présentées avec les critères de performance pour des applications qui utilisent l'estimation de la posture 3D dans quatre configurations (monoculaire, temporelle, multivues et multimodale). L'exactitude est rapportée en MPJPE sur le jeu de données Human3.6M. En gras la meilleure méthode et soulignée pour la deuxième meilleure. Les meilleures approches multimodales utilisant l'IMU (marquées par *) sont évaluées en MPJPE sur TotalCapture pour comparaison.

Méthode avec l'exactitude la plus élevée

Lorsqu'on examine les résultats des méthodes contemporaines, on constate un écart moyen de 10 mm d'erreur entre les méthodes monoculaires et les méthodes multivues. Lorsque les méthodes multivues sont utilisées correctement, la combinaison de la connaissance géométrique de la scène et des optimisations d'apprentissage peuvent aboutir à des erreurs aussi faibles que 20 mm. Toutefois, l'erreur est plus élevée dans un contexte général (ce qui est également vrai pour les méthodes monoculaires). Par exemple, Iskakov et al. (2019) ont obtenu une MPJPE de 34 mm sur un jeu de données différent de celui utilisé pour l'entraînement (malgré 17mm de MPJPE sur le jeu de donnée de validation).

Les IMU sont également un bon moyen d'améliorer la détection multivues. Cependant, pour l'instant, les résultats semblent proches de ceux obtenus avec les données d'image seules, Zhang et al. (2020) obtenant les résultats les plus précis. De futures évaluations sur le jeu de données TotalCapture Trumble et al. (2017) avec des méthodes n'utilisant pas les informations inertielles pourraient aider à la comparaison.

La force et la faiblesse des méthodes multivues viennent du fait qu'elles sont basées sur les paramètres des caméras. Cela contribue à la généralisation du système, car il peut s'adapter à de nouvelles vues de la caméra (voir He et al. (2020)), mais rend également le processus plus complexe car il nécessite une calibration. Une solution est suggérée par Kocabas et al. (2019b), avec un système qui calcule les paramètres de la caméra à la volée avec les articulations détectées comme cibles de calibration. Les

images multiples contiennent beaucoup plus d'informations qu'une seule et les propriétés bien connues de la vision stéréo sont applicables. Il s'avère que la plupart des recherches se sont concentrées sur les images monoculaires et l'estimation 2D, si bien que de nombreuses autres voies restent souvent inexplorées, comme l'exploitation simultanée de données temporelles et multivues.

Les méthodes monoculaires basées sur des séquences vidéo pouvant être traitées hors ligne telles que Cheng et al. (2020) et Pavllo et al. (2019) ont du mérite. Cependant, leur exactitude n'est pas inférieure à 40 mm MPJPE sur Human3.6M, et leur capacité de généralisation n'est pas non plus prouvée, car aucune analyse comparative n'est encore disponible pour elles.

Les méthodes simples mais exactes

Les méthodes monoculaires sont les plus simples, car elles peuvent être alimentées par une simple entrée image ou vidéo. Dans de nombreux cas, elles constituent une nette amélioration par rapport à l'annotation manuelle, car elles ont une exactitude similaire mais fournissent des informations 3D plus riches avec moins de données d'entrée. Lorsque des données vidéo sont disponibles, il est judicieux de favoriser les méthodes basées sur des données temporelles qui fournissent de meilleurs résultats (Cheng et al. (2020), Wang et al. (2020) ou Pavllo et al. (2019)). Kolotouros et al. (2019) proposent également une méthode basée sur des images uniques qui est compétitive avec les techniques basées sur des séquences d'images (MPJPE 41,1mm sur Human3.6M). Cela indique que leur approche consistant à optimiser conjointement un modèle de maillage 3D et à entraîner un réseau de neurone est un moyen efficace de résoudre les problèmes d'estimation de la posture monoculaire.

Bien que la méthode de Wandt and Rosenhahn (2019) n'atteigne pas l'erreur moyenne typique de 40 mm, elle le fait à un faible coût de calcul. Elle constitue donc un bon compromis entre exactitude et rapidité pour l'estimation de la posture en temps réel.

Dans de nombreux cas, un autre facteur important est la flexibilité de la méthode. Pourtant, la plupart des algorithmes n'ont pas été conçus dans cette optique, car ils sont limités aux étiquettes des données d'entraînement. Pour l'étude du mouvement humain, il est courant de devoir définir des points clés personnalisés pour suivre les membres, les segments ou les articulations. Ce problème peut être résolu en utilisant DeepLabCut (Mathis et al., 2018) ou d'autres méthodes qui ne sont pas spécifiques à l'homme comme Zhang and Park (2020). Évidemment, ces méthodes nécessitent tout de même des exemples de vérité du terrain pour chaque nouveau point clé, mais à une échelle beaucoup plus réduite. Une autre possibilité consiste à "fine-tune" des modèles existants avec de nouvelles étiquettes personnalisées.

Conclusion générale sur l'exactitude

Le tableau 3.6 montre que les méthodes les plus précises ont un niveau de robustesse moyen à faible et que peu d'entre elles fonctionnent en temps réel. L'explication vient du fait qu'il s'agit souvent de méthodes complexes qui abordent des problèmes tels que l'occultation avec des modules spécifiques qui augmentent le coût global de calcul. Comme expliqué ci-dessus, les meilleures méthodes sont naturellement des techniques multivues qui exploitent les contraintes géométriques et nécessitent donc des processus de calibrations.

Actuellement, les architectures qui obtiennent la meilleure exactitude sont des algorithmes "top-down" en deux étapes. Ils obtiennent les meilleurs résultats sur une variété de jeux de données de référence dans des configurations d'images monoculaires, de vidéos monoculaires et de vues multiples. Ils s'appuient souvent sur le succès de l'estimation de la posture en 2D, qui est un problème presque résolu (il obtient des scores PCK moyens supérieurs à 90%). De nombreuses approches différentes sont efficaces : régression directe de la 2D à la 3D, initialisation de modèles paramétriques humains, exploitation de séquences temporelles, modules tenant compte des occultations, modèles génératifs, représentation volumétrique des entrées, triangulation multivues, pour ne citer que les plus importantes (voir section 3.4 et fig. 3.2). Un point intéressant est que, logiquement, les méthodes de séquence vidéo sont plus performantes pour les activités présentant une cohérence temporelle, comme la marche ou la course, tandis que les méthodes monoculaires détectent mieux les postures statiques complexes.

Une autre distinction importante peut être faite entre les approches basées sur un modèle et les approches sans modèle. Alors que l'état de l'art en matière de détection 2D n'utilise pas de modèles humains, la détection 3D le fait avec succès. Elles se basent soit sur les "Pictorial Structure" (Belagiannis et al., 2014), soit sur un modèle de maillages 3D comme SMPL (Loper et al., 2015) ou SCAPE (Anguelov et al., 2005). De récentes propositions utilisent également la détection 2D de réseaux de neurones comme initialisation pour leurs modèles avec succès.

Cependant, les techniques sans modèle sont aussi performantes, voire meilleures dans certains cas. L'erreur de détection pour la tâche d'estimation de la posture humaine en 3D a diminué de près de 70 mm au cours de la dernière décennie, principalement grâce aux différentes architectures de réseaux convolutifs. Il est intéressant de noter que ces améliorations sont principalement dues à de meilleurs modules 2D, ce qui peut indiquer que la recherche sur la nature 3D du problème est encore une voie prometteuse.

Robustesse				
Méthode	Type	Exactitude	Niveau de robustesse	Temps réeel
Mehta et al. (2020)	Monoculaire	63.6	élevé	✓
Xu et al. (2019)	Monoculaire	82.4	élevé	✓
Hossain and Little (2018)	Temporelle	58.5	élevé	✓
Cheng et al. (2020)	Temporelle	40.1	moyen	✗
Liu et al. (2020)	Temporelle	45.1	moyen	✗
Iskakov et al. (2019)	Multivues	17.7	moyen	✗
von Marcard et al. (2018)	Multimodal	26.0*	élevé	✗

TABLE 3.7 : Les méthodes les plus robustes. Le niveau de robustesse est défini comme suit : 1 - 3 pré-requis : niveau élevé, 3 - 4 pré-requis : niveau moyen et 5 pré-requis ou plus : niveau faible.

3.5.2 Robustesse

La robustesse est généralement évaluée par les changements de l'exactitude lors de l'évaluation croisée sur différents jeux de données. Moeslund and Granum (2001) proposent d'exprimer la robustesse comme le nombre de conditions pré-requises pour qu'une configuration de capture de mouvement soit opérationnelle. Ils définissent vingt hypothèses relatives au protocole, à l'environnement d'acquisition ou encore à l'apparence du sujet. Certaines d'entre elles ne sont plus nécessaires dans toutes les méthodes actuelles (sujet portant des vêtements spécifiques ou fonds monochromes statiques), mais d'autres sont encore très débattues (occultations, personne seule ou pas de mouvement de caméra). Voici les hypothèses qui restent dans l'état de l'art :

Contraintes de mouvement : Pas de mouvement de la caméra, pas de mouvement rapide du sujet, pas d'occultation (pas d'occultation forte, pas d'auto-occultations). Pour les méthodes temporelles, le nombre d'images et la fréquence d'acquisition peuvent également constituer une limite.

Contraintes liées à l'environnement : Matériel supplémentaire (IMU, scans laser), caméras multiples (avec ou sans paramètres de caméra), etc.

Contraintes du sujet : Une seule personne, la première posture connue, un mouvement parallèle au plan de la caméra (pour certaines approches basées sur des modèles).

Fonds, éclairage et vêtements

La plupart des contraintes d'apparence ne sont plus nécessaires, car les réseaux de convolutifs identifient mieux les invariants visuels significatifs que les anciens extrac-

Méthode	#Frames	Causal
Hossain and Little (2018)	5	✓
Cai et al. (2019)	7	✗
Pavllo et al. (2019)	243	✓
Cheng et al. (2019)	256	✗
Cheng et al. (2020)	128	✗
Liu et al. (2020)	243	✓
Wang et al. (2020)	96	✗

TABLE 3.8 : Pré-requis relatifs aux méthodes temporelles monoculaires. Nombre d'images nécessaires pour obtenir l'exactitude optimale et possibilité d'adapter le système pour qu'il n'utilise que les images du passé (pour une utilisation en temps réel).

teurs de caractéristiques. Cependant, il est difficile d'évaluer la généralisation à des situations réelles, principalement parce que de grands jeux de données de capture de mouvement sont souvent encore capturés dans des studios intérieurs. L'augmentation des données peut fournir de nouvelles scènes grâce à l'extraction de l'arrière-plan ou à la modification de l'éclairage. Avec la sortie de nouvelles solutions commerciales sans marqueur, de nouveaux jeux de données de référence tels que MPI-INF-3DHP Mehta et al. (2016) incluent également de plus en plus de données en condition réelles et en extérieur.

Matériel additionnel & Systèmes calibrés

Deux contraintes sont toujours d'actualité : premièrement, la nécessité d'un matériel spécifique, deuxièmement, la nécessité d'une calibration dans les configurations à vues multiples. Le matériel supplémentaire intervient lorsque la méthode utilise d'autres modalités que les images comme les unités de mouvement inertielles (IMU) et les capteurs de profondeur tels que les caméras infrarouges ou à temps de vol. Les unités de mouvement inertielles fournissent des informations supplémentaires sur l'orientation des membres, mais leurs résultats sont sujets à une dérive après une courte utilisation. Elles sont également plus pratiques que les marqueurs réfléchissants et les combinaisons de capture de mouvement, mais restent intrusives pour le sujet. Différents capteurs de profondeur ont également été utilisés pour déduire la profondeur des articulations directement ou pour construire des caractéristiques plus complexes en tant qu'étape de pré-traitement de l'estimation de la posture. Moins fréquemment, certaines méthodes utilisent le scan 3D de chaque sujet qui est capturé pour ajuster les paramètres de forme des modèles de corps humain. Logiquement, alors que la plupart des recherches sont menées sur des méthodes monoculaires, elles sont toujours surpassées de 10 à 20 mm de MPJPE par les techniques multivues.

Méthode	Calibration	IMU	#Views
Trumble et al. (2017)	✓	✓	4
Qiu et al. (2019)	✓		4
Kocabas et al. (2019b)			4
Iskakov et al. (2019)	✓		2
He et al. (2020)	✓		3
Huang et al. (2019)	✓	✓	4(8)
Zhang et al. (2020)	✓	✓	4(8)

TABLE 3.9 : Pré-requis liés au matériel des modèles à vues multiples. Le nombre de vues nécessaires pour obtenir moins que l'erreur MPJPE de base de 40 mm (meilleurs résultats des méthodes monoculaires) sur Human3.6M est également indiqué (et sur TotalCapture entre parenthèses).

Dans les systèmes multivues, l'étape de calibration est un pré-requis fréquent. Les méthodes multivues qui n'utilisent pas de calibration ou utilisent une calibration partielle ont besoin de plus de vues pour atteindre une exactitude acceptable, tandis que d'autres peuvent produire de bons résultats avec moins de caméras mais ont besoin de paramètres extrinsèques (voir tableau 3.9).

Les pré-requis importants qui sont encore utilisés pour de nombreux systèmes d'estimation de la posture humaine en 3D sont liées au mouvement de la caméra ou du sujet et au protocole d'acquisition. L'un d'entre eux est la limitation à une seule personne dans l'image. Comme les approches "top-down" sont les plus populaires, de nombreuses méthodes supposent ou même prennent en entrée une seule personne. C'est pourquoi certains auteurs recommandent l'utilisation de détecteurs de personnes de l'état de l'art pour recadrer la zone de l'image contenant les sujets individuels. De cette manière, les algorithmes d'estimation de la posture peuvent être appliqués à chaque zone individuellement. Cependant, cette idée doit être adaptée à des configurations temporelles et multivues pour suivre chaque sujet différent. La principale limite réside dans le fait que les membres des sujets se chevauchent dans l'image et ne peuvent donc pas être distinguées par les détecteurs de personnes. Certaines recherches ont abordé cette question en 2D, mais elle reste un problème ouvert pour la 3D avec peu de méthodes spécifiques (Mehta et al. (2018), Mehta et al. (2020)).

Restriction de mouvement

Les anciens systèmes de suivi du mouvement sans marqueur étaient parfois contraints de ne suivre que les mouvements lents de quelques membres pour effectuer une bonne

détection. C'est de moins en moins le cas, mais il est toujours difficile de prédire les mouvements rapides (par exemple, dans les vidéos de sport). Cheng et al. (2020) suggèrent que les données temporelles et spatiales à différentes résolutions peuvent être une solution à ce problème. Un nouveau pré-requis peut être ajouté pour les méthodes basées sur des données temporelles : si la séquence vidéo n'est pas assez longue pour fournir suffisamment d'informations, cela peut être un problème pour l'inférence en temps réel et même pour l'exactitude. En outre, les méthodes temporelles utilisent souvent des informations dans les images futures de la vidéo, ce qui n'est pas adapté au temps réel. Le tableau 3.8 montre que ces méthodes sont plus performantes avec des champs réceptifs temporels variables. Certaines méthodes n'ont besoin que de quelques images passées et futures, tandis que l'exactitude d'autres saturent à plus de 200. Ces méthodes peuvent être adaptées à des clips vidéo plus courts et à une application en temps réel en utilisant uniquement des images passées, mais au prix d'une diminution de l'exactitude. Une autre contrainte forte est l'utilisation de vidéos provenant de caméras en mouvement, mais ce problème est rarement abordé (von Marcard et al., 2018). De nombreuses applications peuvent fonctionner avec l'hypothèse de caméras fixes, mais il existe une quantité importante de données vidéo produites avec des systèmes de coordonnées de caméras mobiles (par exemple, dans les sports de plein air).

Occultations

Des solutions récentes permettent de prédire les postures même en présence de faibles occultations, mais cela reste l'un des principaux défis d'une approche monoculaire. Dans la plupart des scénarios du monde réel ou multipersonnes, il est nécessaire d'y remédier. À cette fin, de nombreuses optimisations au moment de l'entraînement sont utilisées : augmentation des données ou modules spécifiques à l'occultation (Cheng et al. (2020), Cheng et al. (2019), Huang et al. (2019)). Une autre solution pourrait être de considérer la scène 3D autour du sujet. Hassan et al. (2019) montrent que la combinaison de la reconstruction 3D de la scène intérieure et des modèles volumétriques du corps humain peut aider à surmonter les problèmes d'occultation. Leur méthode "Proximal Relationships with Object eXclusion" (relations proximales avec exclusion d'objets) applique des contraintes physiques telles que les contacts avec un environnement statique.

Généralisation

Bien que les modèles de réseaux neuronaux soient des algorithmes orientés donnée, peu d'analyses sont effectuées sur la généralisation à de nouveaux contextes et à des situations naturelles. Les protocoles pour les jeux de données de référence traitent de la généralisation à différents sujets, mais la scène de fond et les actions changent rarement. Les modèles multivues se généralisent le mieux, avec de bonnes performances

en validation croisée entre les jeux de données, probablement en raison des informations géométriques fournies par les matrices de projection des caméras qu'ils utilisent souvent. Cependant, de nouveaux jeux de données de référence sont nécessaires pour évaluer correctement la généralisation. Les recherches futures fourniront probablement des jeux de données naturalistes soigneusement conçus avec de nouvelles solutions d'étiquetage, ou des jeux de données augmentés avec de nouvelles techniques de traitement d'image, et éventuellement des jeux de référence composés de postures synthétiques, pour mettre au défi les méthodes futures.

Conclusion générale sur la robustesse

Le tableau 3.7 présente les méthodes les moins contraintes et leurs performances. La robustesse est liée au nombre de pré-requis (moins il y en a, plus la robustesse est importante). Pour les techniques monoculaires, les méthodes multipersonnes entraînées sur des ensembles de données complexes contenant des occlusions sévères imposent logiquement moins de contraintes. Pour les techniques temporelles, celles qui ne nécessitent pas d'images futures pour l'inférence sont les plus robustes. Hossain and Little (2018) obtient une exactitude maximale avec le moins d'images possible. Cheng et al. (2020) traitent les mouvements rapides et les occlusions mais nécessitent une fréquence d'acquisition plus élevée et un champ réceptif temporel plus large pour produire de meilleurs résultats. La méthode multivues la plus robuste est Iskakov et al. (2019) car elle ne nécessite aucun matériel spécial et peut fonctionner avec seulement deux caméras tout en atteignant une exactitude acceptable (voir tableau 3.9). Enfin, von Marcard et al. (2018) peut effectuer une estimation de la posture de plusieurs personnes dans la nature avec des caméras mobiles. Il faut tout de même noter que les faibles contraintes en termes d'environnement et de sujet se font souvent au prix de fortes contraintes en termes de matériel supplémentaire (IMU et scans).

Tout au long de cet état de l'art, force est de constater que l'obtention de la plus grande exactitude est la principale préoccupation des nouvelles méthodes. Nous avons également constaté que la robustesse du système est régulièrement négligée lors de l'évaluation des algorithmes et que la complexité de la nouvelle technique est rarement prise en compte, en particulier pour certains systèmes complets qui dépasse le stade de prototypes. L'accent général mis sur la plus grande exactitude semble quelque peu trompeur et la recherche de systèmes minimaux "suffisamment exacts" en fonction du domaine d'application pourrait être au moins aussi importante. Par exemple, une très grande exactitude n'est pas pertinente pour la surveillance, mais la robustesse et le fonctionnement en temps réel le sont. Pour les professionnels de la santé qui surveillent le rétablissement des patients, la robustesse est primordiale, à condition que l'exactitude soit suffisante (c'est-à-dire comparable à celle de l'œil humain), mais le fonctionnement en temps réel est seulement souhaitable.

Vitesse

Méthode	Exactitude	Niveau de robustesse	Vitesse	GPU
Mehta et al. (2017)	80.5	moyen	30fps	N/A
Wandt and Rosenhahn (2019)	50.9	moyen	20 000fps	Nvidia Titan X
Xu et al. (2019)	82.4	élevé	120fps	N/A
Mathis et al. (2018)	-	moyen	30fps	Nvidia 1080Ti
Mehta et al. (2020)	63.6	élevé	30fps	N/A
Hossain and Little (2018)	58.5	élevé	300fps	Nvidia Titan X
Liu et al. (2020)	45.1	moyen	3000fps	Nvidia Titan RTX
Pavllo et al. (2019)	46.8	moyen	150 000fps	Nvidia GP100
Trumble et al. (2017)	87.3	faible	25fps	N/A
Huang et al. (2019)	37.5	faible	25fps	Nvidia 1080Ti

TABLE 3.10 : Méthodes fonctionnant en temps réel. Outre les autres critères de comparaison, la vitesse (en inférences sur une image par seconde) et la carte graphique utilisée sont indiquées. Il faut noter que des méthodes telles que Hossain and Little (2018) et Wandt and Rosenhahn (2019) ne prennent pas en compte la vitesse du détecteur 2D dans la première étape de leurs techniques lorsqu'ils renseignent les fps.

3.5.3 Vitesse

Il est plus facile d'évaluer les performances en termes de vitesse en observant simplement la complexité ou le nombre d'opérations uniques nécessaires pour une méthode donnée (en utilisant le nombre d'opérations en virgule flottante par secondes"*FLOPs*"). Lorsque cela est possible, savoir à quelle fréquence d'images une inférence peut être faite est également intéressant pour les applications en temps réel telles que le contrôle, la surveillance ou la réalité virtuelle. Dans ces cas, la principale contrainte est la latence de l'ensemble du système, qui ne doit pas dépasser les limites spécifiées.

Temps-réel

Toutes les méthodes ne sont pas des systèmes complets de capture de mouvement comme Huang et al. (2019) ou Mehta et al. (2017) qui rapportent 25 et 30 images par seconde pour la détection. D'autres ne fournissent que le temps d'inférence par image, sans le tester dans un scénario d'acquisition réel, comme Hossain and Little (2018) et Martinez et al. (2017) qui rapportent chacun 3ms par image. Les architectures les plus profondes ne sont souvent pas adaptées aux applications en temps réel, car une passe à travers le réseau prend trop de temps avec des machines aux configurations courantes. Pour avoir un aperçu de la complexité de ces modèles, on peut regarder le nombre de paramètres qu'ils apprennent (voir tableau 3.11).

Temps d'apprentissage

La plupart des méthodes examinées peuvent être entraînées pour s'adapter à de nouveaux contextes difficiles ou être "fine-tuned" sur de nouvelles données. Il est également important d'avoir une idée du temps d'apprentissage lorsqu'une application envisage un apprentissage en ligne. Le nombre de paramètres est une bonne indication de la profondeur de l'architecture (voir tableau 3.11), qui peut également être retrouvé en utilisant leurs réseaux de détection 2D de base dans de nombreux cas.

Méthode	#Parameters (in million)
Martinez et al. (2017)	4-5
Sun et al. (2018) Hourglass	26
Sun et al. (2018) Resnet50	26
Sun et al. (2018) Resnet101	45
Pavllo et al. (2019)	16.95
Qiu et al. (2019)	560
Iskakov et al. (2019) algebraic	80
Iskakov et al. (2019) volumetric	81
He et al. (2020)	69

TABLE 3.11 : Nombre de paramètres rapporté pour les différentes techniques décrites

Conclusion générale sur la vitesse

Dans la partie vitesse du tableau 3.10, nous signalons toutes les méthodes qui peuvent s'exécuter en temps réel. La vitesse d'inférence est exprimée en images par seconde et le GPU utilisé est également spécifié. La robustesse est variable dans cette collection de méthodes, et l'exactitude semble un peu en dessous de la moyenne pour les méthodes les plus rapides. La plus précise est Huang et al. (2019), qui atteint 37,5 MPJPE sur Human3.6M sans utiliser son composant IMU (elle est également performante sur TotalCapture avec l'ajout des données IMU).

3.5.4 Recommandations pour les utilisateurs

Après cette analyse en fonction des critères de performance, voici nos suggestions pour les différents types d'application :

Interface homme-machine : pour cette catégorie d'applications, la priorité est une forte exactitude et des performances en temps réel. La robustesse dépend de si le système fonctionne dans un environnement prédéfini ou naturel. Avec ces spécifications, les méthodes temporelles pouvant fonctionner en temps réel telles que Pavllo et al. (2019) ou Liu et al. (2020) correspondent le mieux. Hossain and Little (2018) nécessite le calcul de moins d'images mais son exactitude est moindre. D'autres méthodes plus robustes pourraient être des méthodes multipersonnes, comme Mehta et al. (2020), mais les applications de contrôle n'interagissent généralement qu'avec un seul sujet. Enfin, Huang et al. (2019) est la méthode la plus précise fonctionnant en temps réel, mais elle nécessite plusieurs vues.

Sécurité : Les systèmes de surveillance traditionnels doivent généralement être plus robustes pour fonctionner dans des environnements et des sujets variés et changeants du monde réel. La rapidité est également importante car les infractions doivent être identifiées en temps réel. Enfin, une précision importante est moins utile car c'est l'ensemble de la sémantique du mouvement qui est importante pour identifier les actions du sujet. Pour des méthodes fonctionnant en temps réel avec la plus grande robustesse possible, considérez les méthodes monoculaires (Xu et al. (2019)) et temporelles (Hossain and Little (2018)). Les méthodes proposées par Mehta et al. (2020) permettent une robustesse encore plus grande avec une détection multipersonnes en temps réel, si nécessaire. Des méthodes moins robustes peuvent également être envisagées pour obtenir une plus grande exactitude (par exemple, Pavllo et al. (2019) ou Liu et al. (2020)) ou une plus grande rapidité (par exemple, Wandt and Rosenhahn (2019)).

Analyse du mouvement : ces applications fonctionnent souvent dans un environnement contrôlé en laboratoire et les calculs sont effectués hors ligne. L'exactitude est le critère le plus important, avec moins de considération pour le temps réel ou une grande robustesse. Cependant, la performance humaine capturée dans des scénarios de la vie réelle et le besoin de techniques moins intrusives (pour le diagnostic médical ou la rééducation) exigent une meilleure robustesse que les méthodes classiques basées sur les marqueurs (Colyer et al. (2018)). Les configurations multivues produisent désormais des résultats avec une erreur moyenne par articulation inférieure à 30 mm (Iskakov et al. (2019), He et al. (2015)). Lorsque l'utilisation de plusieurs caméras ou la calibration n'est pas possibles, des méthodes temporelles monoculaires telles que Kolotouros et al. (2019) ou Cheng et al. (2020) peuvent être envisagées. Enfin, si la simplicité est un critère, des systèmes faciles à utiliser et flexibles tels que Mathis and Mathis (2019) ou Martinez et al. (2017) produisent des résultats efficients.

Divertissement : La capture de mouvement pour l'animation ou la réalité virtuelle se fait généralement dans un environnement contrôlé, la robustesse n'est donc pas la principale préoccupation. L'exactitude est importante car les postures et les mouvements générés doivent être réalistes (voir 3.2.1 pour les métriques structurelles et perceptuelles). Enfin, la vitesse est moins importante pour le traitement hors ligne qui génère des avatars ou des personnages animés, alors que le temps réel est nécessaire pour les contrôles dans les jeux vidéo ou la réalité virtuelle. Dans le premier cas, les systèmes sans marqueur multivues peuvent remplacer les systèmes classiques de capture de mouvement à moindre coût (par exemple Iskakov et al. (2019) ou He et al. (2020)). Dans le second cas, le fonctionnement en temps réel est possible avec Huang et al. (2019), qui fournit également une grande exactitude dans les configurations multivues (et même plus si l'ajout d'IMUs au sujet est possible).

3.5.5 Défis futurs pour la recherche

De nombreux problèmes concernant l'estimation de la posture humaine en 3D n'ont toujours pas été résolus. Le plus discuté est que l'exactitude est encore insuffisante pour certaines applications (par exemple, dans l'analyse du mouvement). Actuellement, de nombreuses approches monoculaires et temporelles différentes permettent d'obtenir une erreur moyenne de position articulaire d'environ 40 mm, tandis que les approches multivues permettent d'obtenir environ 20-30 mm. Cela entraîne des défis importants :

Pour les techniques *monoculaires*, l'obstacle à leur utilisation généralisée est la cohérence de leur détection dans l'espace et le temps. Les différences d'exactitude entre les articulations doivent être prises en compte et la cohérence temporelle (réduction des instabilités de prédiction) des mouvements doit être mieux respectée. Atteindre 90% d'articulations détectées avec précision, comme le font les estimateurs de posture 2D, est un objectif pour les années à venir (le 3DPCK moyen atteint environ 85% avec des évaluations sur MPI-INF-3DHP).

Pour les configurations *multivues*, le problème est qu'elles présentent des résultats proches de ceux des meilleures méthodes monoculaires malgré l'accès à la géométrie épipolaire et aux informations 3D. La combinaison des progrès récents de l'apprentissage profond (par exemple, réseau récurrent, attention, etc.) et de solides informations préalables sur les structures de la scène provenant de la calibration des caméras pourrait permettre de franchir une nouvelle étape. Une autre piste pour des applications dans des environnements plus contrôlés pourrait être l'utilisation d'autres modalités telles que l'IMU ou les capteurs de profondeur. La fusion d'informations multimodales pourrait alors conduire à des résultats encore meilleurs.

Des performances en *temps réel* sont obtenues par de nombreuses méthodes, mais en utilisant des cartes graphiques puissantes. Le portage de l'estimation en temps réel sur un équipement commercial moyen reste un défi. Actuellement, la plupart des propositions reposent sur des calculs à plusieurs étapes, mais les recherches futures pourraient s'appuyer sur des détecteurs d'objets "single-shot" (par exemple, Liu et al. (2016) (SSD), Redmon et al. (2016)). (YOLO)) pour produire plus rapidement des résultats raisonnables. De même, la recherche sur l'estimation en temps réel de la posture humaine en 3D avec des caméras multivues se développe. Des applications telles que la réalité virtuelle pourrait bénéficier de tels systèmes pour le contrôle basé sur les gestes.

Enfin, il existe des hypothèses fortes sur les occultations et les configurations multi-personnes qui ne permettent pas encore d'appliquer l'estimation de la posture à toute vidéo ou image. Des ensembles de données complexes dans la nature, comprenant parfois des postures difficiles (générées ou capturées en mouvement), commencent à émerger, ce qui suggère qu'il s'agit de questions de recherche prometteuses.

Deuxième partie

Contributions

Chapitre 4

Fusion de données multivues

Contents

4.1 Introduction

La fusion de données est souvent utilisée en traitement d'image et vision par ordinateur pour de nombreuses tâches. Elle consiste à la combinaison des données issues de capteurs ou traitements différents afin d'obtenir une information plus riche ou plus précise que si une seule source avait été utilisée.

Par exemple, des descripteurs d'images représentant leurs différentes caractéristiques sont construits avant d'être encodés et utilisés en entrée de classifieurs. Pour prendre l'exemple de la classification d'espèces dans des images de fleurs, les caractéristiques de forme (*SIFT, HOG*) et de couleurs sont souvent utilisés conjointement Nilsback and Zisserman (2008), Seeland et al. (2017). Différents types de classifieurs (SVM, Random Forest, etc.) sont utilisés, parfois indépendamment pour chaque type de donnée. Avec la popularisation des réseaux de neurones convolutifs, la fusion de données est toujours utilisée et se décline aussi sur les caractéristiques intermédiaires produites par ces architectures.

Plus récemment, des tâches traitant les mêmes données que l'estimation de posture (images de personnes), ont utilisé la fusion de caractéristiques avec succès. C'est le cas, par exemple, de la détection et de la reconnaissance de visages Lu et al. (2017), Ding and Tao (2015) ou encore de la reconnaissance d'actions Simonyan and Zisserman (2014), Boulahia et al. (2021). Ces approches utilisent maintenant pour la plupart des réseaux de neurones convolutifs comme extracteurs de caractéristiques à minima, voir aussi comme le classifieur final.

Néanmoins, le point commun de ces méthodes est toujours l'utilisation avec succès de diverses méthodes de combinaison des caractéristiques. Celles-ci sont produites soit en amont par un pré-traitement des données à l'entrée des modèles de prédiction, soit à la sortie de réseaux de neurones aux architectures spécifiques. Par exemple Boulahia et al. (2021) proposent des méthodes de reconnaissance d'action à l'aide des caractéristiques multimodales, des valeurs de pixel de l'image, de cartes de profondeurs issues de capteurs infrarouges et du squelette des articulations provenant de la capture du mouvement. Les auteurs de Boulahia et al. (2021) proposent trois stratégies de fusion de ces caractéristiques à différentes étapes de leur système.

Il est en effet possible de classer différents types de fusion de caractéristiques fonction de l'étape où ils interviennent dans la prédiction. On parle de "*early fusion*" ou fusion au niveau des données d'entrée si les données issues de différents capteurs ou modalités sont combinées et traitées par un unique modèle. La "*fusion tardive*" utilise les caractéristiques des différentes sources de données en les passant à des modèles spécifiques et la fusion se fait au niveau du processus de prédiction final. Enfin, la fusion dite "*intermédiaire*" est un principe qui consiste à fusionner les caractéristiques intermédiaires produites par des réseaux de neurones pendant l'apprentissage dans une couche de fusion (voir fig. 4.1).

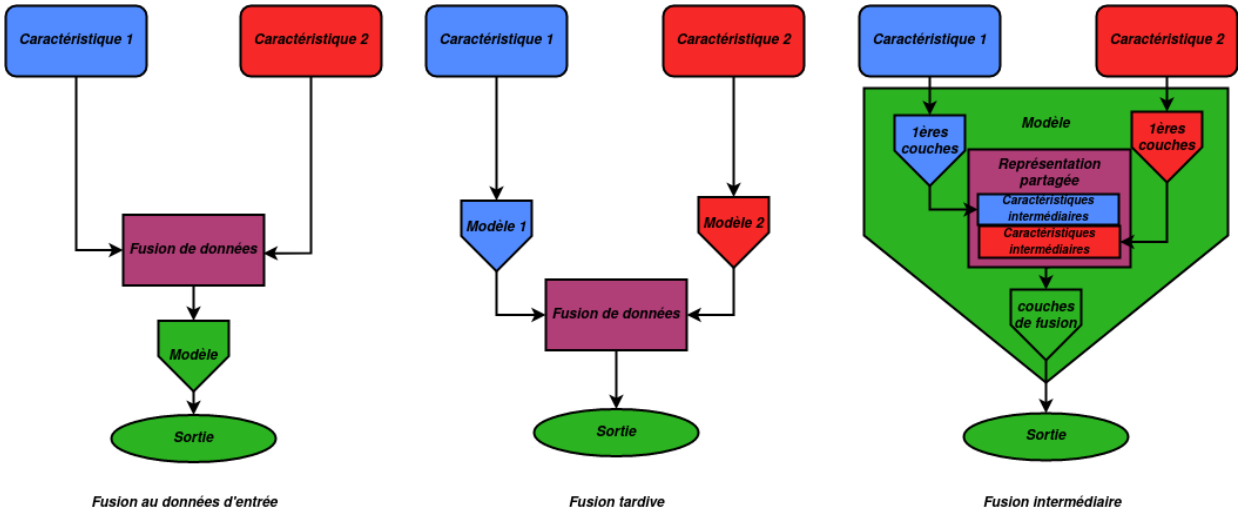


FIGURE 4.1 : Différentes stratégies de fusion de caractéristiques. Les données sont issues de capteurs différents ou de multiples capteurs identiques.

L'estimation de posture est une tâche qui se prête à la fusion de caractéristiques, car les données étudiées sont complexes et variées. En effet, les nombreuses modalités qu'offre la capture du mouvement, que ce soit avec une caméra, dans les installations multivue ou encore celles utilisant des capteurs de profondeurs ou des centrales inertielles (IMU)... Par exemple, Trumble et al. (2017) fusionnent des données 3D reconstruites à partir d'images multivue vidéo et des données d'orientation des articulations provenant d'IMU.

Il est aussi possible dans le cadre d'un système à plusieurs caméras calibrées, sans autres capteurs spécifiques, d'utiliser les modalités qu'offre la 3D. Par exemple, Qiu et al. (2019) utilisent uniquement la modalité apportée par le multivue en prenant en compte l'information supplémentaire donnée par la géométrie épipolaire. Lors de la fusion des caractéristiques, sont produites de meilleures cartes de chaleurs pour locali-

sation de chaque articulation. Une autre possibilité est l'utilisation de représentations intermédiaires volumétriques comme les *"Probabilistic visual hulls"* de Trumble et al. (2017) ou des cartes de localisation tridimensionnelles. Dans ces deux cas, ces représentations sont construites à partir de pré-traitements via des modèles d'apprentissage automatique ou des méthodes traitement d'image.

Malgré la réussite de certaines de ces méthodes dans l'état de l'art, il existe aussi des méthodes de régression de posture directe qui n'utilisent qu'une seule source de donnée. De plus, la plupart des méthodes n'utilisent pas toutes les modalités offertes par un système multivue calibré conjointement aux avancées actuelles en apprentissage sur les données séquentielles (RNN, LSTM : voir sous-section 3.4.2). La motivation des expériences menées sur la fusion de donnée 3D en multivue est donc l'obtention d'une méthode utilisant pleinement les connaissances de la géométrie 3D de la scène ainsi que sur la forme et le mouvement du corps humain. La question soulevée est la suivante : **Est-il possible de tirer parti des informations sur la scène et le sujet en 3D à partir d'images multivue en tant que modalités dans le cadre de fusion de donnée pour l'apprentissage de la posture humaine ?**

La première expérience menée concerne la fusion des trajectoires d'articulation 2D pour la prédiction de posture 3D (4.2). La deuxième consiste en la fusion de silhouettes 2D dans les différentes vues (4.3).

4.2 Fusion de trajectoire de posture 2D

Dans un premier temps, une configuration minimale a été testée. C'est-à-dire l'utilisation des trajectoires 2D provenant des images de chaque vue en tant que sources de donnée indépendante. Pour mettre en place une telle configuration, il a été décidé s'inspirer de la méthode de référence de Martinez et al. (2017) dont l'architecture du modèle et permet le traitement de coordonnées 2D directement. L'objectif étant de pouvoir tester l'influence du type de fusion effectué sur les données multivues. Cette architecture effectue la prédiction de la posture 3D directement à partir de trajectoires 2D, ce qui permet une adaptation facile au cas multivue. En effet, il est assez naturel d'effectuer la fusion des données après la prédiction par l'architecture d'origine ou encore d'adapter l'entrée du réseau en combinant plusieurs trajectoires (respectivement pour la fusion tardive ou pour la fusion aux données d'entrée).

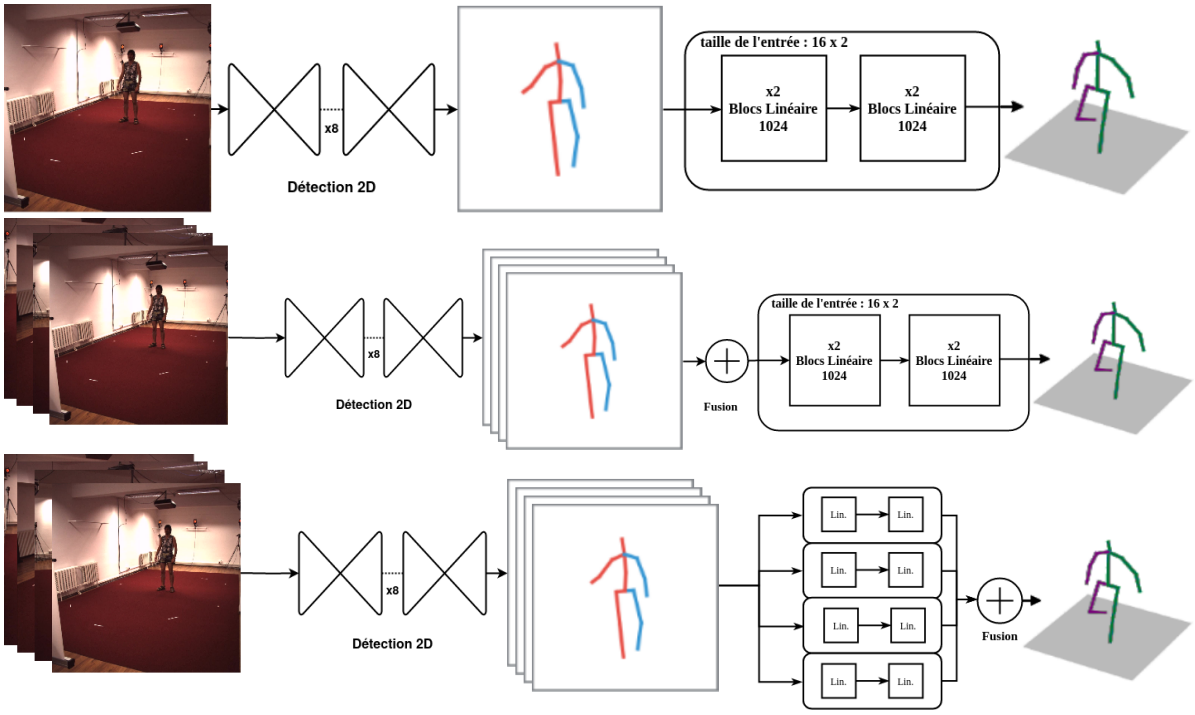


FIGURE 4.2 : Architectures de fusion de donnée multivue. **En haut**, architecture d'origine prenant en entrée une image (monoculaire) Martinez et al. (2017). **Au milieu** et **en bas** adaptation de l'architecture pour prendre des données multivue soit en entrée du réseau, soit en fusionnant l'information à la sortie respectivement. Le détecteur 2D utilisé est "Stacked Hourglass Networks" de Newell et al. (2016).

4.2.1 L'architecture

L'architecture d'origine de Martinez et al. (2017) prends en entrée les coordonnées 2D des articulations produites par le détecteur "stacked hourglass network" Newell et al. (2016) (voir 3.3.3). Le vecteur de données est ensuite directement passé à travers un réseau de neurones enchaînant deux couches linéaires entrecoupées de *normalisation par batch* et de *dropout* (voir section 2.4). Ce processus est répété deux fois avant d'atteindre une couche de sortie linéaire produisant des coordonnées d'articulations en 3D (voir figure 4.2 haut).

La même procédure d'apprentissage est utilisée que dans le papier d'origine (division en ensemble de donnée d'apprentissage sur les sujets 1,5,6,7 et 8 et ensemble de validation sur les sujets 9 et 11, procédure standard sur Human3.6M).

Les données d'entraînement utilisées par les auteurs de Martinez et al. (2017) proviennent du jeu de données Human3.6M Ionescu et al. (2014). Ce jeu de données contient des images multivue et les vérités terrain correspondantes provenant d'un système de capture du mouvement classique (avec marqueurs). Dans un premier temps, les tests reprennent l'adaptation de l'architecture au cas multivue avec la même architecture tout en modifiant la taille des données d'entrée (fig 4.2 milieu). Les données d'entrées de l'architecture d'origine constituent des vecteurs de taille 16 x 2 correspondants coordonnées 2D des 16 articulations et la sortie du réseau donne un vecteur 16 x 3 pour les coordonnées en trois dimensions. Notre méthode adaptée au multivue prend simplement en entrée un vecteur 16 x 2 x 4 pour les quatre vues disponibles dans le jeu de donnée Human3.6M. Ensuite, il est possible de comparer ce résultat avec la fusion après prédiction de chaque posture individuellement (fig 4.2 bas).

4.2.2 Fusion des données d'entrée : présentation des résultats

Plusieurs configurations ont été testées pour la fusion de données multivue des coordonnées 2D avec l'architecture de référence (voir fig 4.3). Tout d'abord, lors de l'entraînement, deux possibilités ont été testées. Soit à partir de trajectoires issues du détecteur "stacked hourglass" Newell et al. (2016), soit à partir des vérités terrain issues de la capture du mouvement re-projeté en deux dimensions. Pour les données de test, la même séparation a été faite uniquement dans le cas où l'entraînement avait également été réalisé avec les trajectoires de la MoCap. L'objectif, pour cette configuration, est de connaître l'exactitude maximale atteignable avec la méthode. C'est le cas théorique dans lequel les coordonnées les plus exactes possibles sont obtenues afin de mieux analyser l'influence du détecteur 2D utilisé sur la performance du système.

Enfin, l'évaluation est conduite selon deux protocoles utilisés dans l'article d'origine : Martinez et al. (2017) et couramment utilisées dans la littérature :

- **Protocole #1** : Pas d'alignement avant le calcul de l'erreur.
- **Protocole #2** : Alignement Procruste aux vérités terrain avant le calcul de l'erreur

Dans les deux cas, les prédictions centrée autour d'une "articulation racine" (approximativement située au niveau du pelvis). Ces protocoles sont utilisés dans le cas où obtenir la position absolue de la posture n'est pas directement possible. Pour l'usage final de ce type de méthode, un détecteur permettant d'obtenir la position de l'articulation racine est nécessaire. L'erreur reportée dans cette section suivant la métrique "MPJPE" (voir sous-section 3.2.1). L'alignement dans le **protocole #2** nécessite la recherche d'une rotation et d'une translation optimale entre la posture prédite et la vérité terrain. L'algorithme de Kabsch-Umeyama (Kabsch (1976) et Umeyama (1991)) est couramment utilisé pour les obtenir.

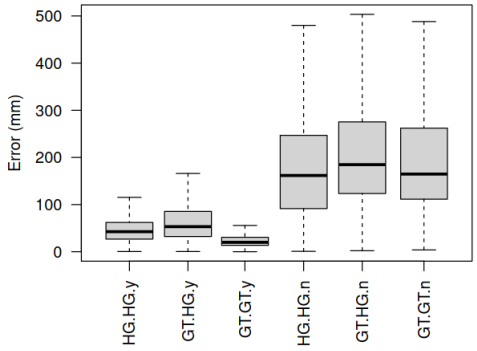


FIGURE 4.3 : Répartition de l'erreur (MPJPE) lors du test de fusion de données multivue. **entraînement.validation.alignement**. **entraînement** : HG="Hourglass detection", GT= Ground Truth. **validation** : HG="Hourglass detection", GT= Ground Truth. **alignement** : n=pas d'alignement, y=alignement rigide au pelvis

En analysant les résultats obtenus sur les différentes configurations ainsi que pour la méthode d'origine (fig. 4.3 et tab. 4.1) voilà ce qui est observé :

- Pas de gain d'exactitude et forte dispersion de l'erreur pour le protocole #1 (voir fig. 4.4)
- Amélioration de l'exactitude dans le protocole #2(voir fig. 4.5)
- Marge d'amélioration importante par rapport à l'utilisation des données de capture du mouvement.

Pour le protocole #1 (voir fig. 4.4) l'exactitude obtenue avec la fusion multivue est bien plus faible qu'en monoculaire sans fusion de données (plus de 100mm de différence avec la méthode d'origine). Il semblerait, en isolant cette observation dans cette configuration, que la fusion des trajectoires n'apporte pas d'informations supplémentaires utiles à la prédiction en trois dimensions, voir qu'elle en dégrade les résultats. Si l'on observe les résultats obtenus pour chaque action plus en détails, les actions les plus "difficiles" pour la méthode de Martinez et al. (2017) (comme "sitting down") restent celles qui sont le moins bien détecté avec la fusion multivue. Cela vient confirmer que même dans des cas où beaucoup d'auto-occultation se produisent, la fusion multivue n'apporte pas les caractéristiques suffisantes pour caractériser le sujet en 3D avec ce type de fusion de données en entrée.

Cependant, la configuration utilisant des coordonnées réalignées semble produire une erreur moyenne légèrement plus faible qu'avec une seule vue (à noter que cette erreur totale est calculée en excluant l'articulation du pelvis, ce qui ne semble pas être le cas dans le code source de la méthode d'origine accentuant d'autant plus l'écart entre

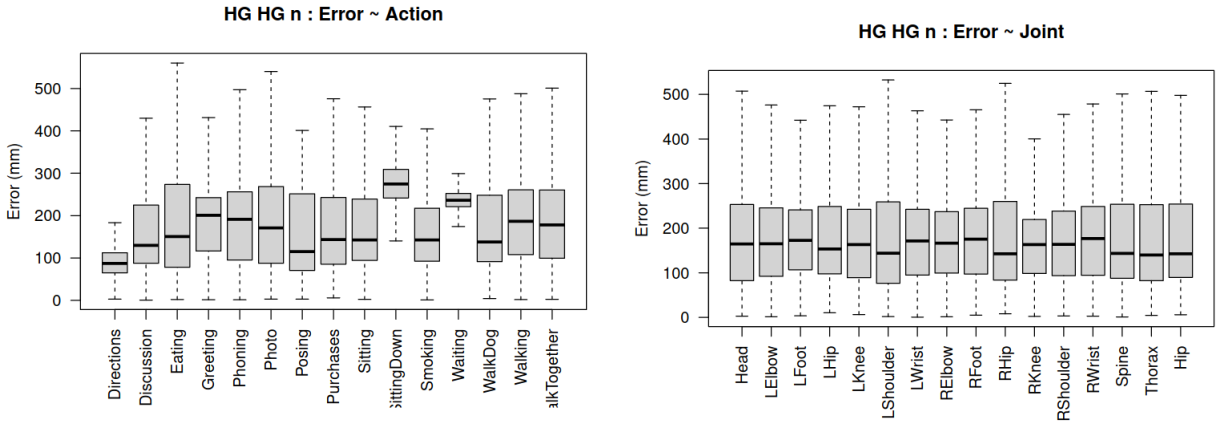


FIGURE 4.4 : Comparaison par actions et par articulations de l'erreur de la méthode de fusion multivue (protocole #1 : pas d'alignement au pelvis)

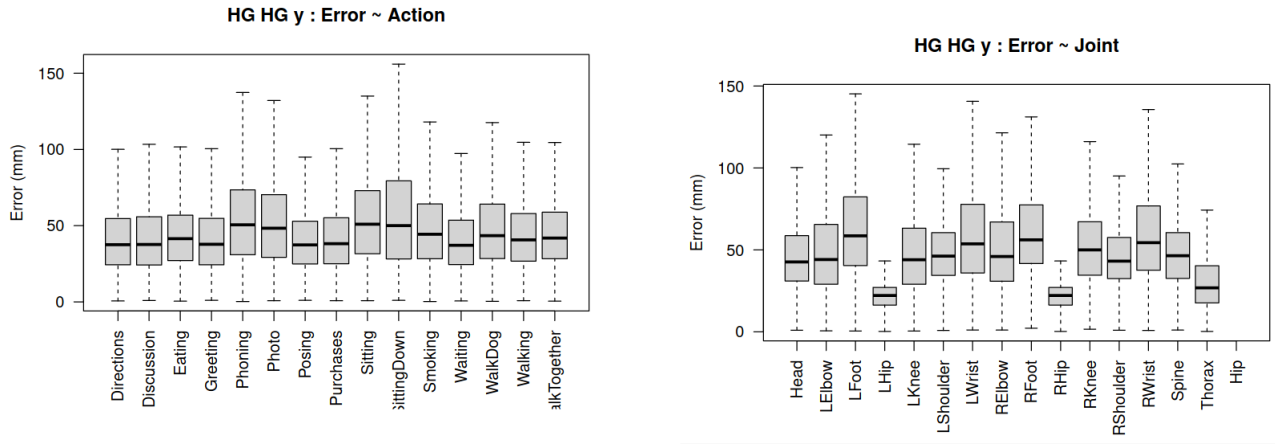


FIGURE 4.5 : Comparaison par actions et par articulations de l'erreur de la méthode de fusion multivue (protocole #2 : alignement au pelvis)

Entraînement/Test	Protocole #1	Protocole #2
Fusion multivue GT/GT	-	*24.7*
Martinez et al. (2017) (GT/HG)	-	60.52
Martinez et al. (2017) (HG/HG)	**67.5**	52.5
Fusion multivue GT/HG	200.2	67.5
Fusion multivue HG/HG	171.8	**50.5**

TABLE 4.1 : Erreurs moyennes des différentes configurations de fusion de données multivue en entrée suivant l'entraînement du modèle avec la vérité terrain (GT) ou le détecteur Hourglass Networks (HG) (Newell et al., 2016). Protocole #1 : pas d'alignement. Protocole #2 : alignement. La première ligne donne un ordre d'idée de la valeur théorique que pourrait atteindre le système si les détections 2D étaient "parfaites" (en utilisant la vérité terrain lors de l'évaluation)

les deux méthodes). La dispersion de l'erreur est logiquement beaucoup plus faible et diminue plus l'articulation est éloignée du point de la taille du sujet (*Hip*). Dans ce cas-là, les actions "difficiles" en monoculaire décrites plus haut sont bien mieux détectée avec le multivue ("sitting down" et "sitting" qui en moyenne sont détectées avec 65mm à 75mm d'erreur dans la méthode d'origine le sont autour de 50mm avec la fusion de donnée). Les caractéristiques multivue sont hypothétiquement mieux apprises dans une configuration plus simple où les coordonnées relatives sont utilisées en entrée d'un réseau.

Le gain d'exactitude est néanmoins assez faible et il est intéressant de noter qu'il y a un écart important entre les 52,5 - 50,5mm (monoculaire et multivue respectivement) et les 24,7mm d'erreur moyenne obtenue lorsque les reprojections des coordonnées de la capture de mouvement sont utilisées. Ce résultat peut donner une idée de la précision maximale que pourrait atteindre une méthode similaire qui utiliserait des prédictions 2D "sans erreurs". Utiliser des trajectoires 2D plus précises avec des détecteurs récents est alors une perspective d'amélioration. Il faut tout de même noter que, comme cela a souvent été décris pour d'autres tâches, si le modèle utilise des trajectoires produites par le même détecteur lors de l'apprentissage et du test, l'exactitude est plus forte que si on utilise les vérités terrain pourtant plus exactes en 2D (voir fig. 4.3, *HG.HG.y* contre *GT.HG.y*).

4.2.3 Comparaison avec la fusion tardive

Il existe plusieurs manières de fusionner les résultats produits par des modèles en sortie (voir tableau 4.2. Tome et al. (2018) proposent une simple moyenne des coordonnées 3D issues de Martinez et al. (2017) alors que Pavlakos et al. (2017b) utilisent

Méthode	Protocole #1	Protocole #2
Martinez et al. Multivue (Tome et al., 2018)	57.0	-
Pavlakos et al. (2017b)	56.9	-
Fusion en entrée (notre méthode)	171.8	**50.5**

TABLE 4.2 : Comparaison avec les méthodes de fusion tardive. Martinez Multivue est une méthode de base proposée dans Tome et al. (2018) qui consiste simplement à moyenner les postures 3D produites par la méthode de Martinez et al. (2017) dans chaque vue. Pavlakos et al. (2017b) utilise le modèle "Pictorial Structure" (voir 3.3) pour la fusion.

un modèle 3D de déformation du corps humain en 3D. Ce modèle est alimenté par des cartes de chaleurs 2D issues de la localisation des articulations dans chaque vue. Ici aussi, on peut constater que la fusion en entrée est dépassé par ces deux méthodes dans le protocole #1 sans alignement. Cependant, avec le protocole #2, notre produit des résultats comparable à l'état de l'art, mais pour des méthodes monoculaires. L'erreur 3D obtenue est proche de celles de la méthode Pavlakos et al. (2017a) utilisant le modèle de corps humain 3D (*"pictorial structures"* voir section 2.3). Ces résultats ne sont pas suffisants pour constituer le détecteur utilisé pour le prototype, car moins précis que des méthodes multivues existantes. La section suivante explore une autre approche utilisant le même type d'architecture appliquée à des données de silhouettes plutôt que des coordonnées d'articulations.

4.3 Fusion de masques de silhouettes

Pour poursuivre la fusion de données multivue des tests ont été effectués sur la combinaison d'images binaires représentant le masque de la silhouette des personnes dans chaque vue. Les caractéristiques utilisées en entrée du réseau étant plus riches que des coordonnées 2D comme dans l'architecture testée précédemment. Pour obtenir des silhouettes précises, et ce, d'une manière suffisamment rapide pour l'inférence en temps réel, une méthode de segmentation de personne originale a été conçue.

4.3.1 Segmentation de silhouette temps-réel

La méthode YOLACT (Bolya et al., 2019) réalise la segmentation d'objet dans des images. Elle permet la segmentation d'instance sur les 80 catégories d'objets du jeu de données MS-COCO (Lin et al., 2015). Cette méthode a la particularité, contrairement à la plupart de l'état de l'art en segmentation sémantique, de produire des résultats en temps réel.

Méthode	mask AP	FPS	Classes
Yolact (Bolya et al., 2019)	30.0	24	MS-COCO (80)
Mask-RCNN (He et al., 2018)	36.1	8.6	MS-COCO (80)
Mask-RCNN (He et al., 2018)	45.1	8.6	masques personne
Pose2Seg (Zhang et al., 2019)	55.5	-	masques personne
Yolact-Densepose	44.0	24	masques personne

TABLE 4.3 : Comparaison des segmentations de silhouette. mask AP = précision moyenne pour les IOU (intersection over union) par pixel de 0.5 à 0.95.

C'est le cas de Mask-RCNN (He et al., 2018) qui atteint une AP (voir section 3.2.1) de 45.1 pour les masques de personne, mais avec seulement 8.6 images par secondes. Pose2Seg (Zhang et al., 2019) est une des rares méthodes spécialisée pour la segmentation de personne (contrairement aux deux autres méthodes entraînées sur tout le jeu de donnée MS-COCO (Lin et al., 2015) Elle obtient le plus haut score d'AP (55.5) mais n'a pas été testé pour l'inférence en temps réel.

Il a donc été décidé de se baser sur l'architecture de YOLACT en le ré-entraînant sur un jeu de donnée spécialisé. Le principe de YOLACT est Ce jeu de donnée est un sous ensemble de MS-COCO (Lin et al., 2015) pour l'estimation dense de la position des personnes dans les images : DensePose (Güler et al., 2018). Ce jeu de donnée vise à une association des pixels à une position sur surface du corps humain. Pour la segmentation de silhouette, seuls les masques de chaque personne fournis sont utilisés.

La stratégie adoptée est le fine-tuning de YOLACT en utilisant le modèle pré-entraîné sur le jeu de donnée MS-COCO entier. L'idée est que le modèle d'origine est déjà entraîné en partie sur des images contenant des humains. Les jeux de données sont suffisamment proches et bien labellisés pour pouvoir entraîner à nouveau directement le réseau sur Densepose.

Le résultat obtenu fournit une AP plus importante (44.0) que le modèle d'origine (bien que sur une seule classe) tout en conservant sa vitesse d'inférence (24 images par secondes).

4.3.2 Fusion de silhouette multivue

Après avoir obtenu les silhouettes dans chaque vue sur le jeu de données Human3.6M plusieurs tests ont été effectués en reprenant l'architecture présentée dans 4.2. L'objectif est de tester la fusion en entrée avec les silhouettes de chaque vue plutôt que les trajectoires 2D.

Dans un premier temps, les seules caractéristiques utilisées sont les masques de silhouette de chaque vue. Ceux-ci ré-échantillonnés et normalisés avant d'être passé en entrée à l'architecture de base précédemment décrite. Les masques de pixels correspondant à la silhouette (4 x 1000 x 1000 pixels) son sous-échantillonnés par valeur moyenne (*Average Pooling*) avec un noyeau de taille 8x8 et un pas de 8 générant une taille d'entrée de 4 x 125 x 125. Cette méthode simple n'a pas fourni des résultats de posture exploitables. En ajoutant la posture à la silhouette et en suivant le même processus, une erreur moyenne (MPJPE) de 116.2mm est obtenue, ce qui n'est pas un résultat probant.

La fusion de silhouette et de posture multivue en entrée n'améliore pas le résultat pour ce type d'architecture de réseau avec de multiples couches linéaires (contrairement à la version utilisant uniquement les postures multivues qui semble l'améliorer dans certains cas). Utiliser la silhouette pour la fusion tardive pourrait être une piste de recherche à l'avenir. Une solution pourrait être de l'utiliser en sortie des modèles pour régulariser la posture 2D issue de chaque vue avant la prédiction 3D. Une autre possibilité serait d'utiliser les masques pour paramétrer un modèle de contrainte du corps humain comme Pavlakos et al. (2017b). Enfin, une architecture de fusion intermédiaire qui est déjà utilisé dans certains cas en estimation de posture pourrait être testée dans de futures recherches. Ces possibilités sont abordées dans la section suivante.

4.4 Conclusion

4.4.1 Fusion intermédiaire

Une possibilité pour exploiter les données d'images multivues pour l'estimation de posture serait d'utiliser la fusion intermédiaire (voir 4.1). Cette manière de combiner les données de plusieurs sources au cours de l'apprentissage de réseaux de convolution a déjà été exploitée avec succès dans d'autres domaines. Pour l'estimation de posture, c'est surtout dans le cadre de l'utilisation de modalités supplémentaires (principalement les IMU) que l'on trouve des exemples Trumble et al. (2017)Huang et al. (2019). Le processus mis en place va souvent d'abord extraire une caractéristique intermédiaire de chaque modalité avant de les fusionner dans une couche de représentation partagée. Ces modèles utilisent des représentations volumétriques de la scène à partir du multivue : (*probabilistic visual hulls* pour Trumble et al. (2017) et volumes multicanaux pour Huang et al. (2019)). Ce sont des exemples de fusion de données en entrée produisant des caractéristiques plus complexes que celles des expériences précédentes. Ce type de représentation peut être une piste pour la création de nouvelles architectures utilisant uniquement des données provenant d'images. Cependant, il existe très peu de méthodes qui n'utilisent que la géométrie connue des caméras dans la scène les unes par rapport aux autres avec des données purement visuelles pour la fusion de donnée. Le meilleur exemple d'une telle méthode est la "learnable triangulation of human pose" de Iskakov et al. (2019). Dans leur méthode volumétrique (deuxième méthode présentée dans l'article) les auteurs produisent des cartes de chaleur des articulations dans chaque vue et les re-projettent dans une grille de voxels formant des volumes agrégés de la posture. Cette "représentation partagée" est ensuite utilisée en entrée d'un modèle d'affinement de volume en posture Moon et al. (2018). C'est l'une des seules méthodes qui utilise un processus de fusion intermédiaire avec des images multivues.

4.4.2 Pistes de recherche

Il semblerait que la fusion de données multivue est prometteuse en utilisant des modèles d'apprentissage profond avec des représentations intermédiaires utilisant la géométrie connue de la scène. L'un des enjeux des futures recherches dans ce domaine est la conception de ces architectures et des modules qui les composent. Trumble et al. (2017) utilisent une succession de convolution 3D et de max pooling. Huang et al. (2019) utilisent le même principe en adaptant le réseau en *"stacked hourglass network"* de Newell et al. (2016) avec des convolutions 3D pour produire des cartes de chaleurs volumétriques. Enfin, Iskakov et al. (2019) utilisent le modèle de Moon et al. (2018) qui est une architecture d'auto-encodeur qui transforme une posture sous forme de volume en plusieurs cartes d'une chaleur volumétriques par articulation.

De plus, si les vidéos multivues sont accessibles, il est possible d'intégrer la nature séquentielle du mouvement dans l'apprentissage des modèles. C'est déjà le cas pour Trumble et al. (2017) qui font passer les caractéristiques intermédiaires issues de l'IMU et du multivue à travers une LSTM avant la fusion. Des méthodes plus récentes monoculaires ont démontré l'efficacité des architectures se basant sur des convolutions temporelles (*Temporal Convolutional Network* ou TCN) Pavllo et al. (2019)Cheng et al. (2019)Cheng et al. (2020). Une piste de recherche pourrait être l'adaptation au cas multivue de ce type de méthode en utilisant des convolutions 3D sur des représentations plus riches telles que décrites plus haut.

Enfin, un défi important pour les méthodes multivues est lié à la problématique de la calibration et la diversité des angles de vue possibles à chaque acquisition. En effet, si des méthodes basées sur l'apprentissage utilisent un jeu de donnée dont les vues sont similaires pour toutes les séquences, cela peut poser un problème de généralisation. Une solution peut être d'utiliser la géométrie des caméras directement pour la construction de la représentation partagée comme c'est le cas dans Iskakov et al. (2019). Cependant, la formation de jeu de données contenant des angles de vue plus diversifiés et complexes pourrait permettre une meilleure caractérisation des postures dans tous les contextes.

4.4.3 Synthèse et limite

Les modèles entraînés avec la fusion en entrée de postures multivues sur des architectures de base inspirées de Martinez et al. (2017) ont produit des résultats compétitifs pour des coordonnées 3D absolues. Cependant, l'utilisation de masques de silhouette n'a pas apporté de gain d'exactitude, soulignant le besoin d'une fusion intermédiaire ou de représentation en entrée plus riches. Dans les deux cas, le système possède un faible nombre de paramètres (et utilise une segmentation de silhouette à plus de 24 images par secondes) ce qui permet l'inférence des postures en 3D en temps réel.

Cependant, des méthodes multivues plus complexes et nécessitants un plus grand nombre de contraintes d'acquisition produisent une exactitude plus élevée. Dans la contribution suivante, le prototype du système d'acquisition couplé avec plusieurs détecteurs 2D de l'état de l'art sera testé et évalué. Dans la continuité de ces travaux sur la fusion de données et en utilisant des méthodes de triangulation connues Hartley and Zisserman (2004), la fusion tardive sera donc évaluée dans le cadre du prototypage du système.

Chapitre 5

Prototype

Contents

5.1 Introduction

Ce chapitre est consacré à la description du prototype de système d'estimation de posture humaine en trois dimensions et en mouvement sans marqueur. Le nom qui a été donné à ce prototype est *KeenMT*. Ses caractéristiques impliquent un ensemble d'étapes successives qui permettent de passer de vidéos acquises d'un sujet en mouvement depuis plusieurs angles de vue à une série de coordonnées 3D. Ces coordonnées correspondant à plusieurs points d'intérêt situé sur les membres et articulations et sont suivies tout au long du mouvement. Le prototype intègre des algorithmes d'estimation de postures 2D appliqué à chaque vue (voir chapitre 4), mais permet aussi la calibration, l'acquisition et la triangulation des postures dans la scène en trois dimensions. La nécessité de maîtriser tout le processus de la capture du mouvement répond à plusieurs enjeux scientifiques.

La validation du prototype et son évaluation permettent en effet de répondre à la question de la faisabilité d'un tel système qui est le sujet central de cette thèse. L'objectif est de fournir un comparatif détaillé de la nouvelle méthode proposée avec l'état de l'art et les méthodes *"gold standard"*. Ce comparatif est effectué au regard des deux critères principaux pour des applications scientifiques et médicales ciblées : l'exactitude et la robustesse (voir section 2.2 sur les critères d'évaluation).

La deuxième considération importante concernant la validation in situ du prototype est la comparaison en fonctionnement des différents algorithmes qui le composent. Avoir un prototype fonctionnel permet en effet une comparaison a posteriori des performances obtenues en utilisant différents modules pour chaque étape du processus. C'est principalement le cas pour les estimateurs de pose 2D : les architectures de base de l'état de l'art seront comparées à partir des prédictions finales obtenues après la triangulation en sortie du système.

Selon les critères d'évaluations précédemment cités (voir chapitre 2), le meilleur compromis entre précision, robustesse et prix permettant la capture du mouvement sans marqueurs se base sur des méthodes mêlant analyse de donnée visuelle et connaissance des scènes 3D à partir d'image provenant de multiples vues de la scène.

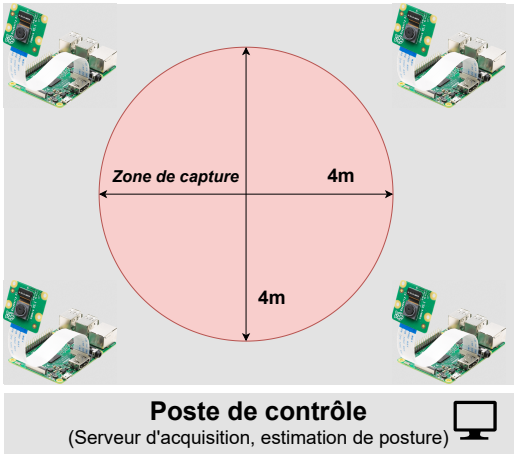


FIGURE 5.1 : Schéma du prototype avec le volume de capture couvert.

5.2 Description du prototype

Le système peut être divisé en étapes de la manière suivante :

1. Acquisition (5.3)
   (a) Calibration des caméras (5.3.1)
   (b) Synchronisation des horloges (5.3.2)
   (c) Capture des vidéos synchronisée (5.3.3)
2. Estimation de posture (5.4)
   (a) Estimation de posture 2D dans chaque vue (5.4.1)
   (b) Triangulation 3D de la posture (5.4.2)
   (éventuellement régularisation des coordonnées)
3. Évaluation de l'erreur (5.5)

L'étape d'acquisition synchronisée avec l'ensemble des caméras du système est décrite dans la section 5.3. Pour fonctionner, cette acquisition nécessite au préalable un processus de synchronisation des flux vidéos capturés ainsi qu'une calibration des différentes caméras pour permettre, plus tard, la reconstruction 3D. La partie du système qui permet l'estimation de la posture est décrite dans la section 5.4. Elle consiste principalement dans la triangulation de coordonnées 3D à partir des point 2D préalablement estimées dans les vues provenant de chaque caméra. Enfin, la dernière étape permet le calcul de l'erreur de détection si une vérité terrain est disponible à partir de la capture du mouvement optique. Cette étape est décrite dans la section 5.5. Il est également possible d'utiliser le système au départ de chacune des trois étapes indépendamment.

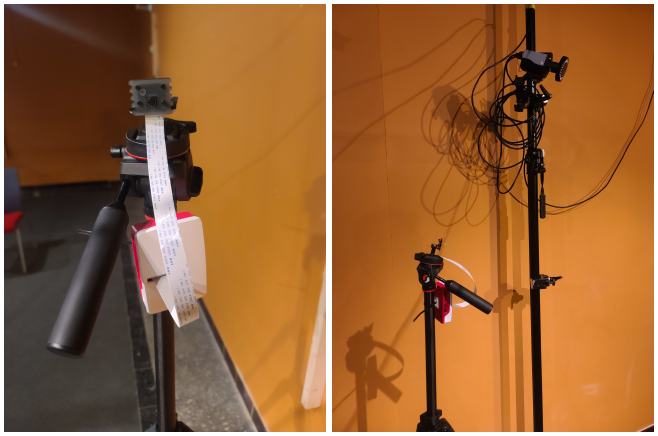


FIGURE 5.2 : Gauche : Une caméra du système avec Raspberry pi. Droite : Module caméra Raspberry et capteur Qualisys.

D'un point de vue matériel, l'installation consiste en quatre micro-ordinateurs SBC (ordinateur à carte unique) de marque Raspberry Pi 4 connectés à des modules caméras sur mesure. Ces périphériques servent de clients d'acquisition et sont connectés par le protocole DHCP sur un serveur d'acquisition et de calcul qui réalise l'estimation de posture (voir figure 5.2). Il est important de noter que cette installation a un but démonstratif et doit permettre de comprendre si ses inférences sont reproductibles dans un cadre plus général avec des périphériques et capteurs standards. C'est pour cette raison que lors des tests d'acquisition et d'estimation, la résolution choisie et de 1000x1000 avec une fréquence d'acquisition de 24 Hz (alors que le module caméra permet une résolution de 1080p à 30 Hz en théorie).

De plus, un tel système utilisé tel quel permet la démonstration de la possibilité d'estimation la posture humaine pour le prix relativement faible comparativement à la capture de mouvement avec marqueurs. L'ensemble du système coûte moins de 600 € et peut être réduit à moins de 400 € en utilisant une version miniaturisée de la carte utilisée (Raspberry Pi Zero ou autre SBC à bas coût supportant un module caméra).

Une telle installation permet ainsi la démonstration du fonctionnement du prototype avec les spécifications nécessaires :

- Sans marqueurs
- Bas coût
- Multivue synchronisé
- Équivalent caméras standard

	KEENMT
Caméras	
Modèle de caméra	Raspberry Pi Module Caméra v2
Nombre de caméras	4
Résolution à l'acquisition	1000x1000
Fréquence d'acquisition	24 Hz
Synchronisation	Logicielle
Connectique	
Client d'acquisition	Raspberry Pi IV
Câblage	Gigabit Ethernet
Liaison poste de ctrl.	Concentrateur Ethernet
Serveur	DHCP

TABLE 5.1 : Matériel utilisé pour tester l'acquisition et le fonctionnement du système.

C'est avec le matériel décrit dans le tableau 5.1 que l'ensemble des tests réalisé en parallèle avec un système de capture de mouvement classique ont été effectués.

5.3 L'acquisition

Dans un premier temps, le système permet l'acquisition synchrone de plusieurs vidéos d'une même scène sous différents angles de capture. Cette capture est possible grâce à la mise en place d'un protocole d'acquisition qui permet la centralisation des données sur une machine non-directement connectée aux capteurs. Ce "poste de contrôle" va permettre de faire le lien avec les caméras et de lancer l'acquisition. Il permet aussi d'effectuer l'ensemble des traitements suivants pour l'extraction de la posture.

Cependant, avant de lancer des acquisitions, une première étape cruciale de calibration des caméras est nécessaire pour pouvoir effectuer la triangulation des articulations dans la scène. Cette étape n'est pas nécessaire si les paramètres des caméras sont connus ou pour des analyses sur des jeux de données existants, Human3.6M (Ionescu et al., 2014) par exemple. Ensuite, une autre étape importante doit être effectuée à chaque démarrage du système. Elle consiste à synchroniser les flux vidéo capturés via chaque caméra. Cette synchronisation est obligatoire afin de permettre une mise en correspondance des points d'articulation dans chaque vue correspondant à des emplacements similaires dans le temps.

5.3.1 Méthode de calibration

Le but de la calibration est d'estimer les paramètres intrinsèques et extrinsèques des caméras afin d'obtenir les informations nécessaires pour la récupération des coordonnées d'articulations dans la scène 3D dans un repère global. Les paramètres intrinsèques correspondent aux caractéristiques optiques de chacune des caméras du système. Les paramètres extrinsèques contiennent les matrices de rotation et vecteurs de translation qui permettent le passage du repère de la scène vers celui des caméras. C'est avec ces paramètres que la triangulation de la posture est possible dans les étapes futures du processus. Si l'acquisition est effectuée en parallèle avec de la capture de mouvement (comme pour la validation du prototype), il est aussi nécessaire de calibrer les caméras vidéos du système avec le système qui sert à détecter les marqueurs pour pouvoir calculer l'erreur de détection dans le même repère.

Une manière possible pour obtenir ces paramètres est la suivante : Utilisation d'une mire de calibration en échiquier dont les points visibles depuis chaque caméra sont ensuite mis en correspondance. La première étape consiste à estimer les paramètres intrinsèques initiaux des caméras avec la méthode de Zhang (2000). Il est ensuite possible de déterminer la pose des caméras en résolvant le système déterminé à partir des points projetés dans la scène :

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

Il est alors possible d'appliquer l'algorithme de Levenberg-Marquardt (Eade, 2013) pour minimiser l'erreur de reprojection (et obtenir la distorsion), c'est-à-dire la somme totale des carrés des distances entre les points caractéristiques observés et leur reprojection en utilisant les estimations actuelles des paramètres des caméras. Voir la section 2.5 pour plus de détails sur la calibration dans un système multivue.

Il existe de nombreuses implémentations pour ces différentes étapes de la calibration, de la capture de mires de formes et tailles différentes à l'algorithme utilisé pour détecter les points mis en correspondance. Pour résumer les étapes d'une calibration du système :

1. Capture d'images de calibration pour chaque caméra (paramètres intrinsèques)
2. Capture d'images de calibration avec toutes les caméras (paramètres extrinsèques)
3. Détection des coins de l'échiquier dans les images de calibration
4. Calculs des paramètres intrinsèques puis extrinsèques

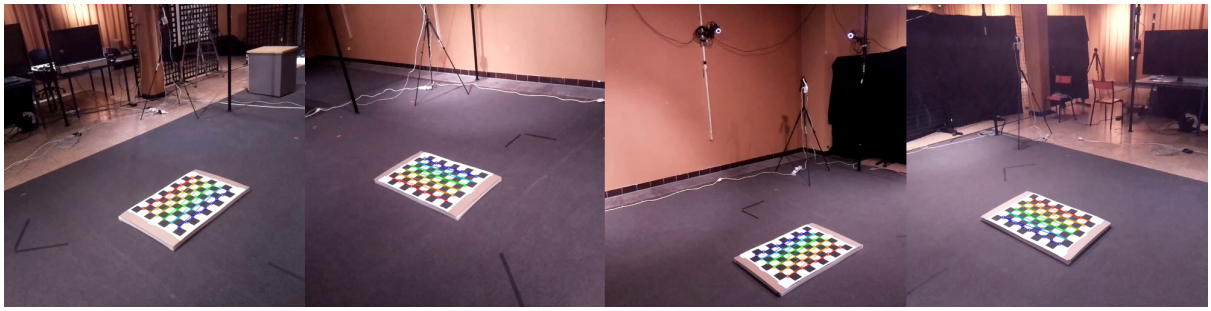


FIGURE 5.3 : Calibration du système à 4 caméras avec échiquier. Images de calibration pour les paramètres extrinsèques obtenues avec l'utilitaire de EasyMocap

Lors de l'évaluation du prototype, les implémentations d'OpenCV (Bradski, 2000) ont été utilisés pour la détection de l'échiquier et le calcul des paramètres des caméras. L'utilitaire de EasyMocap (voir fig. 5.3) a permis de vérifier et corriger si nécessaire les détections des coins de la mire. Le détecteur de coins utilisé se base sur l'opérateur de Förstner Förstner and Gülch (1987) afin de détecter les coins avec une précision sous-pixel sur le motif interne de l'échiquier préalablement localisé dans l'image. Enfin, l'implémentation d'OpenCV de l'algorithme de calibration de Zhang (2000) permet d'obtenir les matrices de paramètres intrinsèques de chaque caméra et les coefficients de distorsion. Il est ensuite possible de calculer les paramètres extrinsèques des caméras à l'aide d'une image de la mire ou celle-ci est visible dans chaque vue (voir fig. 5.3

En pratique, pour pouvoir réussir la calibration avec les caméras choisies, la mire utilisée doit être suffisamment grande pour que l'algorithme de détection (par exemple **?**) puisse extraire les coins de l'échiquier dans la scène. Cette taille dépend de la surface de la zone de capture et des caractéristiques des caméras utilisées (champ de vision et profondeur de champ). La mire utilisée pour la calibration du système pour un volume de capture de 50m² a par exemple été imprimé au format A1 (10x7 carreaux). C'est d'autant plus important pour l'étape de calcul des paramètres extrinsèques où celle-ci doit être entièrement visible par toutes les caméras simultanément. Il peut être nécessaire de vérifier manuellement la bonne détection des coins de l'échiquier à chaque étape de la calibration.

5.3.2 Synchronisation Temporelle

La synchronisation des horloges peut être effectuée avec le protocole NTP en mode client serveur local. Le principe est d'utiliser l'ordinateur du poste de contrôle comme serveur NTP et de l'utiliser pour synchroniser les horloges des cartes d'acquisition.

Peu de temps avant le début de l'acquisition, l'ensemble des micro-ordinateurs Raspberry Pi se synchronisent et réduisent leur décalage au temps de l'horloge du poste de contrôle à un écart de l'ordre du millième de secondes. Au lancement de l'acquisition, un temps de départ est diffusé à toutes les caméras qui vont s'activer et commencer la capture vidéo au temps indiqué dans le message.

Cependant le délai introduit par cette synchronisation logicielle, le système n'est pas adapté à la capture de mouvements rapides. Pour ce type d'utilisation, il faut ajouter un système de synchronisation plus élaboré, mais aussi utiliser des caméras avec une fréquence d'acquisition plus importante, ce qui complexifie le système et le rend moins accessible. Il est également possible d'ajouter un composant électronique pour réception de signal GPS RTC si une synchronisation à d'autres périphériques est nécessaire.

5.3.3 Capture vidéo

L'acquisition vidéo du mouvement d'une personne est possible une fois la calibration des caméras effectuée. Si les caméras sont déplacées, une nouvelle calibration est nécessaire. Elles peuvent être positionnées de part et d'autre de la scène pour couvrir le volume de capture voulu. Il peut être utile de prendre en compte les objets occultant dans la scène lors de la répartition des caméras. Une fois l'installation du matériel effectuée, le système permet l'acquisition vidéo selon le protocole d'acquisition suivant (voir figure 5.4) :

1. Chargement du serveur d'acquisition sur le poste de contrôle et attente de la connexion des clients.

2. Lancement à distance des clients d'acquisition sur les raspberry pi avec les paramètres de la capture nécessaires.

3. Une fois l'ensemble des clients connectés, lancement du départ de la capture depuis le poste de contrôle : diffusion du message de départ à l'ensemble des clients connectés.

4. Réception du message de départ, attente pour le départ, synchronisé de la capture pour chaque client puis capture vidéo.

5. Récupération centralisée des vidéos sur le poste de contrôle.

Lors des essais avec le système de capture du mouvement traditionnel, plusieurs problèmes sont posés : la synchronisation vidéo avec le traçage des marqueurs et la fréquence d'acquisition. Pour les essais et la validation du prototype, une synchronisation matérielle a été mis en place à l'aide d'un flash lumineux repérable dans la scène

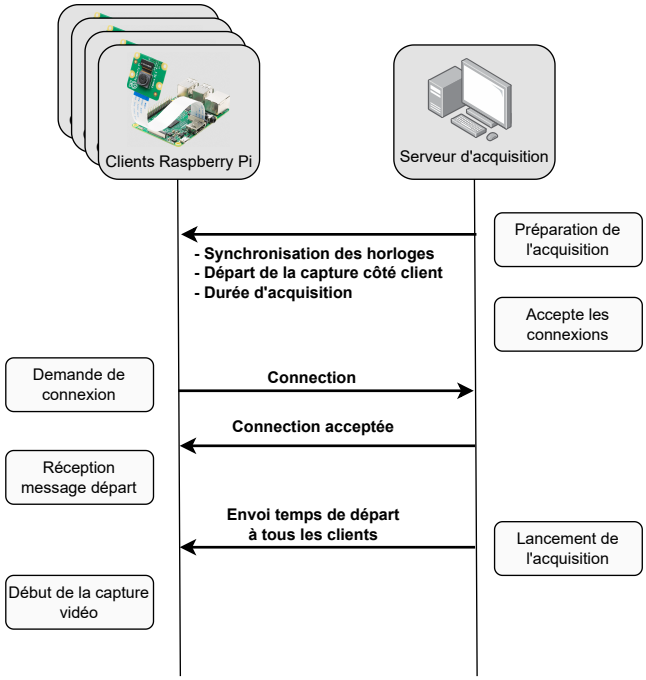


FIGURE 5.4 : Protocole d'acquisition.

et avec les caméras faisant partie du système de capture du mouvement commercial utilisé.

5.4 L'estimation de posture

Après avoir capturé les vidéos et extrait l'ensemble des images multivues de la séquence, il est possible d'estimer la posture du sujet filmé. Ce processus est composé des deux étapes principales suivantes :

- Estimation de posture 2D dans chaque vue.
- Reconstruction de la posture 3D dans la scène.

La première étape d'estimation de posture 2D peut utiliser tout type de détecteurs de l'état de l'art qui fournit une posture sous forme de points d'articulation. Dans la section suivante, l'impact du choix du détecteur sera étudié sur la triangulation finale. Une analyse plus fine en fonction du type de mouvement réalisé et des articulations est aussi importante pour mieux comprendre les performances des différents modèles. Pour permettre cette comparaison, les images utilisées proviennent du jeu de données à grande échelle Human3.6M (Ionescu et al., 2014). Les données qu'il contient offre la

diversité de mouvements nécessaire à l'analyse.

La deuxième étape de l'estimation de la posture et le passage de la 2D à la 3D. Il est effectué via une triangulation dans la scène à partir des paramètres obtenus lors de la calibration. Cette étape permet le passage des coordonnées d'articulation dans le repère image vers le repère de la scène. Il est parfois possible, sur une séquence vidéo, d'obtenir des configurations avec lesquelles les points dans certaines vues sont erronément prédits en 2D. Cela se traduit par des fausses triangulations en dehors du volume de capture, ce qui peut avoir un impact très important sur l'erreur finale pour une quantité de valeurs aberrantes très faible comparativement. Pour cela, il peut être nécessaire d'effectuer une régularisation de ces points en 3D en utilisant la nature séquentielle du mouvement en observant les images suivantes et précédentes dans la vidéo.

5.4.1 Détection 2D

Les résultats présentés dans les tableaux 5.2 et 5.3 sont issus du calcul de l'erreur 3D (MPJPE) après triangulation par rapport à des vérités terrains issues de la capture du mouvement optique. Trois détecteurs de posture 2D sont comparés ici : *"Resnet-152"* He et al. (2015), *"Hourglass networks"* Newell et al. (2016) et *"High resolution networks"* Sun et al. (2019). Ces modèles sont les architectures de réseaux de neurones de base de la majorité des méthodes d'estimation de posture actuelle (voir section 3.3.3). Suivant le protocole couramment employé sur Human3.6M, les modèles utilisés ont été entraînés sur l'ensemble d'entraînement constitué des séquences des sujets 1, 5, 6, 7 et 8 et les séquences des sujets 9 et 11 sont utilisés pour l'évaluation. Les séquences vidéos sont séparées en 15 activités différentes reproduites 2 fois par chaque sujet.

Enfin, aucun alignement procruste n'est effectué par rapport au pelvis comme dans Martinez et al. (2017). La posture triangulée est directement comparée aux coordonnées d'articulation dans le repère 3D de la scène (voir 4.2).

Modèle 2D	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
HgNet. Newell et al. (2016)	53.3	56.3	63.9	54.8	94.4	59.5	55.7	64.3	94.0	249.0	87.8	50.4	58.8	59.6	58.7	73.2
Res-152 He et al. (2015)	57.6	55.7	62.6	52.1	90.0	60.8	56.4	60.7	162.9	**79.9**	94.1	50.5	56.1	68.9	55.2	72.3
HrNet. Sun et al. (2019)	**49.1**	**43.2**	**41.0**	**39.3**	**49.5**	**43.8**	**44.8**	**42.0**	**55.1**	432.9	**44.1**	**36.4**	**44.0**	**46.0**	**38.9**	**59.1**
Filtrage hors vol. capt.																
HgNet. Newell et al.	53.3	56.3	63.9	54.7	91.3	59.5	52.9	62.4	85.3	108.2	75.4	50.4	58.5	58.5	57.9	65.6
Res-152 He et al.	57.2	55.6	62.6	52.1	83.6	60.8	54.9	59.5	107.2	79.9	78.1	50.5	56.1	68.5	54.0	66.0
HrNet. Sun et al.	41.5	35.6	32.8	**32.0**	41.7	36.3	37.2	34.3	41.6	52.0	36.4	28.8	36.5	38.6	31.5	37.0

TABLE 5.2 : Comparaison de l'impact du modèle choisi pour la détection 2D sur les résultats après triangulation sur Human3.6M (en MPJPE). Pas de régularisation des valeurs extrêmes.

L'erreur moyenne pour chaque action directement après triangulation est reportée dans le tableau 5.2. L'architecture *"High resolution networks"* surpasse les deux autres pour la plupart des actions d'en moyenne 6 mm. Les activités les moins bien détectées sont celles qui sont le plus sujettes à contenir des occultations ou confusion inter membres du fait de la présence d'une chaise de bureau dans la scène (comme "téléphoner", "s'assoir" et "fumer"). Il semble néanmoins que le détecteur *"High resolution networks"* soit moins affecté. L'autre problème majeur se situe pour l'action "s'assoir par terre" qui contient un très grand nombre de détections erronées en 2D à cause du phénomène d'auto-occultation (le sujet pratiquement allongé par terre cache une grande partie de ses membres pour plusieurs caméras). Cela cause des problèmes de valeurs aberrantes en dehors du volume de capture, causant une erreur moyenne peu représentative (et ce encore plus distinctement pour *"High resolution networks"*).

Modèle 2D	Direct.	Disc.	Mang.	Sal.	Tel.	Photo	Pose	Ach.	Assis.	S'ass.	Fum.	Att.	Prom.	March.	MarchAv.	Moy.
RANSAC																
HgNet. Newell et al. (2016)	52.7	57.1	63.5	53.2	110.0	63.1	50.9	61.7	86.4	124.6	72.8	51.4	60.6	56.7	57.2	67.5
Res-152 He et al. (2015)	57.8	56.2	63.2	52.3	140.0	66.5	55.5	60.7	344.3	80.3	88.9	50.6	56.8	70.1	54.5	89.9
HrNet. Sun et al. (2019)	40.4	34.9	33.1	32.1	40.4	43.1	36.0	35.13	43.5	72.2	35.5	29.0	35.8	38.2	31.7	38.1
"Least Median of Squares"																
HgNet. Newell et al.	58.9	59.3	58.6	57.1	71.9	61.6	55.5	62.1	72.6	78.9	65.3	56.8	62.1	54.8	54.2	61.8
Res-152 He et al.	54.4	54.0	52.1	53.0	67.1	58.5	52.4	56.2	82.3	67.6	60.6	51.6	55.8	61.4	50.3	58.7
HrNet. Sun et al.	**34.5**	**32.9**	**28.2**	**32.0**	**33.6**	**35.7**	**32.3**	**31.8**	**31.0**	**30.6**	**31.0**	**28.6**	**33.5**	**33.8**	**30.5**	**32.1**

TABLE 5.3 : Comparaison de l'impact du modèle choisi pour la détection 2D sur les résultats après triangulation sur Human3.6M (en MPJPE). Régularisation des valeurs extrêmes.

Une étape de régularisation permet de donner une meilleure idée de la performance de chaque détecteur en lissant ces valeurs à l'aide des prédictions sur les images précédentes dans la vidéo. Les limites du volume de capture étant connues, il est possible d'appliquer les valeurs correctes précédemment obtenues pour les quelques points problématiques avant le calcul de l'erreur. Ces points représentent moins de 300 valeurs aberrantes sur les plus de 65 000 images du sous-ensemble d'évaluation, mais dégradent fortement l'exactitude après l'étape de triangulation du fait de l'utilisation de prédictions 2D erronées mises en correspondance avec des points valides. Ils peuvent être corrigés pendant la triangulation ou en post traitement de plusieurs manières décrites dans la sous-section suivante. Les résultats obtenus après cette étape de régularisation sont reportées dans le tableau 5.3.

Les détecteurs *"Hourglass networks"* et surtout *"High resolution networks"* sont les plus affectés par la régularisation (principalement pour des erreurs causées sur l'action "s'assoir par terre", voir figure 5.5). *"High resolution networks"* ne dépassent notamment plus les 60 mm d'erreur pour aucune action du jeu de données.

Artic.	HgNet	HgNet (vc)	HgNet (lms)	ResNet	ResNet (vc)	ResNet (lms)	HrNet	HrNet (vc)	HrNet (lms)
Taille	31.9	31.9	49.2	**27.1**	**27.1**	40.4	30.4	30.4	32.1
TailleD	52.4	52.4	63.5	49.3	49.3	59.2	**35.9**	**35.9**	37.8
GenouD	58.2	54.0	53.2	58.0	57.6	45.4	34.9	26.8	**23.3**
PiedD	112.6	103.1	85.7	141.6	108.6	81.8	36.5	35.8	**30.3**
TailleG	51.1	51.1	65.9	55.8	55.8	63.4	**42.3**	**42.3**	42.7
GenouG	66.6	66.4	56.1	70.4	70.1	59.2	236.2	39.0	**32.0**
PiedG	148.8	124.6	100.8	201.0	142.3	99.9	46.8	39.9	**32.7**
Thorax	49.0	41.4	47.5	**38.3**	**38.3**	39.1	**38.3**	**38.3**	39.7
Tête	145.2	137.4	120.8	115.3	115.3	114.7	**12.2**	**12.2**	15.7
ÉpauleG	50.1	41.8	45.7	**39.2**	**39.2**	41.4	42.8	42.8	41.4
CoudeG	58.1	55.1	47.8	58.1	58.1	48.6	34.9	34.6	**27.3**
PoignetG	72.0	65.8	47.3	66.9	65.7	44.5	46.3	45.8	**27.8**
ÉpauleD	47.0	44.2	47.2	**42.1**	**42.1**	43.7	44.4	44.3	43.6
CoudeD	76.9	51.3	45.0	55.0	55.0	49.0	44.0	38.0	**27.6**
PoignetD	78.3	62.7	50.9	66.2	66.0	50.1	48.6	48.5	**27.9**

TABLE 5.4 : Erreur par articulation. w/r : avec régularisation des valeurs extrêmes. "HgNet" : Hourglass Networks (Newell et al., 2016), "Resnet" : Resnet-152 (He et al., 2015), "HrNet" : High Resolution Networks (Sun et al., 2019). (vc) : filtrage des valeurs hors du volume de capture, (lms) : least median of square.

Une analyse par articulation pour avoir une comparaison plus fine des détecteurs 2D est présentée dans le tableau 5.4. Ici aussi, il semble que *"High resolution networks"* surpasse les deux autres modèles pour la plupart des articulations en terme d'exactitude. Cette différence est tout particulièrement visible pour les pieds et la tête avec une différence d'erreur de 65 à 160 mm avec les deux autres méthodes. Cependant, *"Resnet-152"* produit de meilleures prédictions en moyenne pour les épaules et le point central de la taille (avec ou sans la régularisation).

Les articulations les plus sujettes à produire des valeurs aberrantes rectifiées après régularisation sont situées dans la partie inférieure du corps (pieds et genoux principalement) et presque jamais dans la partie supérieure (pour la tête, les épaules ou les poignets). Cela peut tendre à confirmer l'hypothèse de l'occultation et auto-occultation, ces membres étant plus saillants lors des actions en question, car moins proches de la chaise ou du sol.

Enfin, le détecteur *"Hourglass networks"* semble bien plus sujet à produire des erreurs aberrantes et il semble que ces erreurs soient faites à travers 12 articulations sur les 15 évaluées contre 9 pour *"High resolution networks"* et 5 pour *"Resnet-152"*. Enfin, il est intéressant de noter qu'en utilisant *"least median of squares"* pour discriminer les vues les moins utiles à la triangulation (voir section suivante) l'exactitude

augmente pour de nombreuses articulations (coudes, poignets, genoux, pieds) mais diminue aussi parfois pour d'autres (taille et tête).

Pour conclure, *"Resnet-152"* semble moins affecté par les occultations (voir figure 5.5) et produit moins d'erreurs dans plusieurs vues, mais *"High resolution networks"* est globalement plus précis à travers toutes les actions et articulations.

5.4.2 Triangulation

Les résultats obtenus pour comparer les méthodes précédentes sont produits à partir de la triangulation de coordonnées dans des images issue de l'estimation de posture 2D et avec les paramètres des caméras obtenus lors de la calibration. Avec ces informations, il est possible d'utiliser une méthode standard de triangulation N-vues telle que la méthode linéaire proposée par dans Hartley and Sturm (1997) Hartley and Zisserman (2004).
Pour n vues et j articulations soit :

- $\mathbf{P}_n$: n matrices de calibration 3×4
- $\mathbf{P}^{iT}$: la $i$ème ligne de $\mathbf{P}$
- $\tilde{\mathbf{y}}_\mathbf{j} = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$: un point dans la scène en 3D
- $\mathbf{y}_\mathbf{j} = \begin{bmatrix} u_n \\ v_n \\ 1 \end{bmatrix}$: ce même point en 2D dans une image

Il est donc possible de déterminer : $\mathbf{y}_\mathbf{j} = \mathbf{P}\tilde{\mathbf{y}}_\mathbf{j}$ au facteur d'échelle k près :

$$k \begin{bmatrix} u_n \\ v_n \\ 1 \end{bmatrix} = \mathbf{P}\tilde{\mathbf{y}}_\mathbf{j}$$

Et donc :

$$ku_n = \mathbf{P}^{1T}\tilde{y}_j, kv_n = \mathbf{P}^{2T}\tilde{y}_j \text{ et } k = \mathbf{P}^{3T}\tilde{y}_j$$

En éliminant k :

$$u_n\mathbf{P}^{3T}\tilde{y}_j - \mathbf{P}^{1T}\tilde{y}_j = 0 \text{ et } v_n\mathbf{P}^{3T}\tilde{y}_j - \mathbf{P}^{2T}\tilde{y}_j = 0$$

Dans le cadre d'une calibration avec n-vues, il est alors possible d'écrire le système suivant à partir de la matrice $\mathbf{A}$ de taille $2n \times 4$:

$$\mathbf{A}\tilde{\mathbf{y}}_\mathbf{j} = 0 \text{ avec } \mathbf{A} = \begin{bmatrix} u_n\mathbf{P}_1^{3T} - \mathbf{P}_1^{1T} \\ v_n\mathbf{P}_1^{3T} - \mathbf{P}_1^{2T} \\ u_n\mathbf{P}_2^{3T} - \mathbf{P}_2^{1T} \\ v_n\mathbf{P}_2^{3T} - \mathbf{P}_2^{2T} \\ \vdots \\ u_n\mathbf{P}_n^{3T} - \mathbf{P}_n^{1T} \\ v_n\mathbf{P}_n^{3T} - \mathbf{P}_n^{2T} \end{bmatrix}$$

Pour résoudre le système, la solution choisie est la méthode linéaire dite "homogène" Hartley and Sturm (1997). Elle consiste à appliquer la décomposition en valeurs singulière à la matrice $\mathbf{A}$. La dernière colonne de la matrice des vecteurs de base de $\mathbf{A}$ (matrice $\mathbf{V}$), corresponds à la solution qui minimise $||A\tilde{y}_j||$ quand $||\tilde{y}_j|| = 1$. Cette solution correspond au vecteur propre de la plus petite valeur propre de $\mathbf{A}^T\mathbf{A}$.

Une fois la triangulation effectuée, il est possible de régulariser les coordonnées pour réduire les erreurs extrêmes comme présenté dans la section précédente. La solution naïve est d'utiliser la connaissance des limites du volume de capture pour filtrer les valeurs aberrantes. La solution consiste à modifier les points d'articulations triangulés en dehors du volume de capture connu en utilisant les mêmes points dans les images voisines de la vidéo. Cependant, ce procédé a de nombreux défauts. Tout d'abord, il est impossible à mettre en place si l'on ne connait pas les bornes du volume de capture dès la calibration des caméras. De plus, il ne permet pas une correction des points mal triangulés à l'intérieur même de la zone de capture. Enfin, elle modifie arbitrairement des valeurs de coordonnées.

Une solution plus robuste serait de s'inspirer de la méthode de *"bundle adjustement"*, souvent employée dans le cadre de la reconstruction en trois dimensions ou de la photogrammétrie. L'idée est de minimiser l'erreur de reprojection en utilisant un algorithme de moindres carrés non-linéaire après la suppression des valeurs aberrantes. Iskakov et al. (2019) proposent une telle approche en appliquant l'algorithme RANSAC (Random Sample Consensus) afin de ne garder que les vues produisant une erreur de reprojection inférieure à seuil prédéfini. Une fonction de coût robuste (Huber Loss) est ensuite utilisée pour calculer l'erreur de reprojection des vues sélectionnées et appliquer les moindres carrés. Une autre solution proposée Hartley and Zisserman (2004) est de remplacer l'algorithme RANSAC par un *"Least Median of Squares* où les combinaisons de vues utilisées ont un score basé sur la médiane des erreurs de reprojection de tous les points sélectionnés.

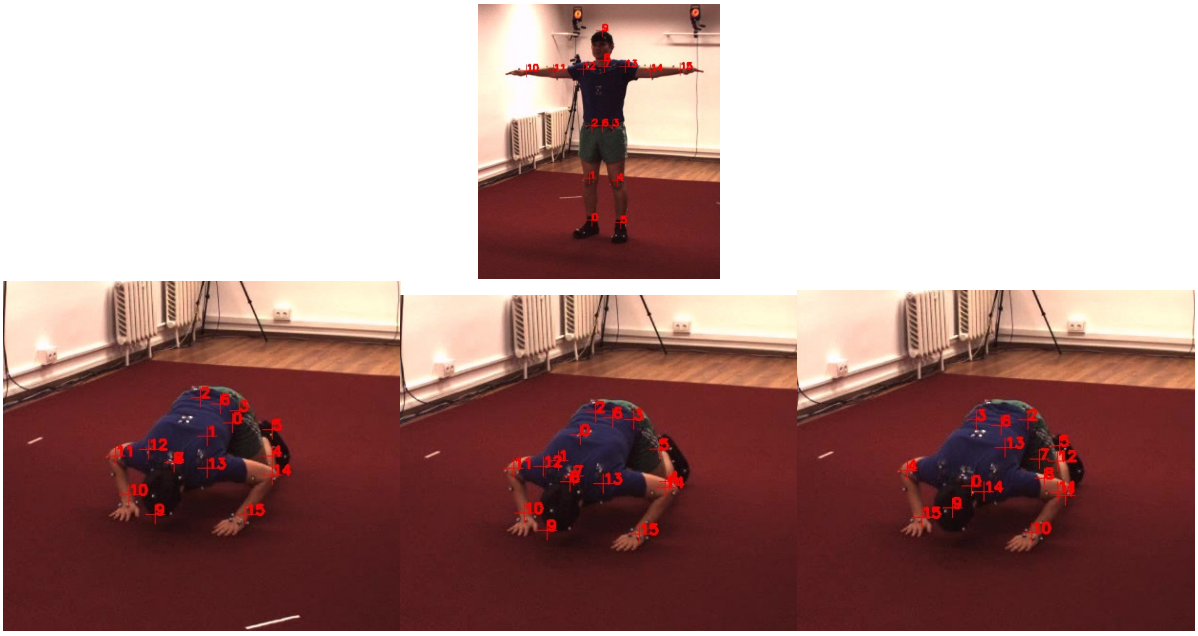


FIGURE 5.5 : Cas d'échec pour détection 2D. Haut : Image de référence. De gauche à droite Hgnet, Resnet, Hrnet.

En utilisant un RANSAC sur 10 itérations par points pour éliminer les vues qui ne contribuent pas à une bonne triangulation à la manière de Iskakov et al. (2019), les reprojections obtenues sont parfois toujours incorrectes. Cela peut s'expliquer par le trop faible nombre de vues utilisé qui peut causer un échec de RANSAC dans le cas où les valeurs aberrantes sont majoritaires. La méthode choisie utilise la médiane des erreurs de reprojection des point dans les quatre vues comme détaillée plus haut. Cependant, au lieu d'échantillonner aléatoirement des couples de vues à chaque itération comme pour la méthode de type RANSAC, l'ensemble des combinaisons de 2 et 3 vues sont testées. Celle possédant le score de reprojection médian dans chacune des quatre vues le plus faible est sélectionnée. Le modèle de départ est initialisé avec la combinaison et le score pour la triangulation utilisant les quatre vues. Cela nécessite autant d'opérations que les 10 itérations prédéfinies de la méthode RANSAC et tend à produire des résultats plus stables sur l'ensemble des données et modèles testés.

5.5 Évaluation du prototype

Enfin, le système permet la quantification de l'erreur de détection produite par l'estimation de posture sous forme de MPJPE (voir 3.2.1). Cette évaluation des performances est possible soit sur le jeu de donnée à grande échelle Human3.6m avec les vérités terrain fournies (voir section précédente), soit à partir de vérités terrain per-

sonnalisées issues de la capture conjointe avec un système de motion capture avec marqueur.

Dans cette section, la validation du fonctionnement du prototype sera effectué sur plusieurs séquences vidéos réalisées dans une plateforme de capture du mouvement avec notre propre protocole. Cette validation permet de tester le prototype directement sur une situation pour lequel il est conçu (occultation avec d'autres objets de la scène, mouvement fréquemment étudiés, etc). De plus, cela permet d'affiner le choix du détecteur 2D et de connaître la capacité de généralisation de ceux-ci en testant leur performance sur des images issues d'un autre jeu de donnée. Enfin, la robustesse du système sera discutée à travers 3 aspects et/ou contraintes : d'abord les contraintes posées par l'installation du système d'un point de vue matériel et logiciel. Ensuite, par rapport à la calibration et l'acquisition. Enfin, grâce à la capacité du système à être utilisé dans des environnements variés.

5.5.1 Protocole d'évaluation

Afin d'évaluer les performances du système in situ comparativement à un système de capture de mouvement avec marqueur du commerce, le protocole suivant a été mis en place. Le prototype décris en 5.2 est installé en parallèle d'un système de capture du mouvement Qualisys équipé de 9 caméras Miqus M3 pour la capture avec marqueurs et 2 caméras Miqus Vidéo synchronisées et calibrées. Les caméras du prototype et de la MoCap sont calibrées indépendamment. Pour vérifier la synchronisation logicielle dans les séquences vidéos de l'acquisition, un signal lumineux est émis dans la scène au départ de l'acquisition. Enfin, l'ensemble des suivis d'articulations et les post-traitements sont réalisés via le logiciel constructeur (*"Qualisys Track Manager"*, voir figure 5.6).

Pour le choix des articulations suivies avec des marqueurs, un ensemble de 12 points prédéfinis par le système ont été choisis (parmi les 42 marqueurs du *"sports markerset"* proposés par Qualisys). Certaines articulations comme les points de la taille où des coudes sont calculés à partir du point central du segment entre 2 marqueurs placés sur le sujet. Ce choix correspond aux articulations labellisées dans le jeu de données MS-COCO Lin et al. (2015) (à l'exception de cinq points du visage : les yeux, le nez et les oreilles) avec lequel de nombreux modèles sont entraînés (voir fig 5.7).

Les séquences capturées concernent une action répétée 3 fois par 2 sujets différents. Cette action est proche d'un test utilisé en science du mouvement humain pour étudier le mouvement des membres supérieurs : le *"reaching"* ou *"functional reach test"* Duncan et al. (1990)Thompson and Medley (2007)Faity et al. (2022). L'action demandée est de saisir un objet posé sur une table basse devant laquelle celui-ci est assis sur une

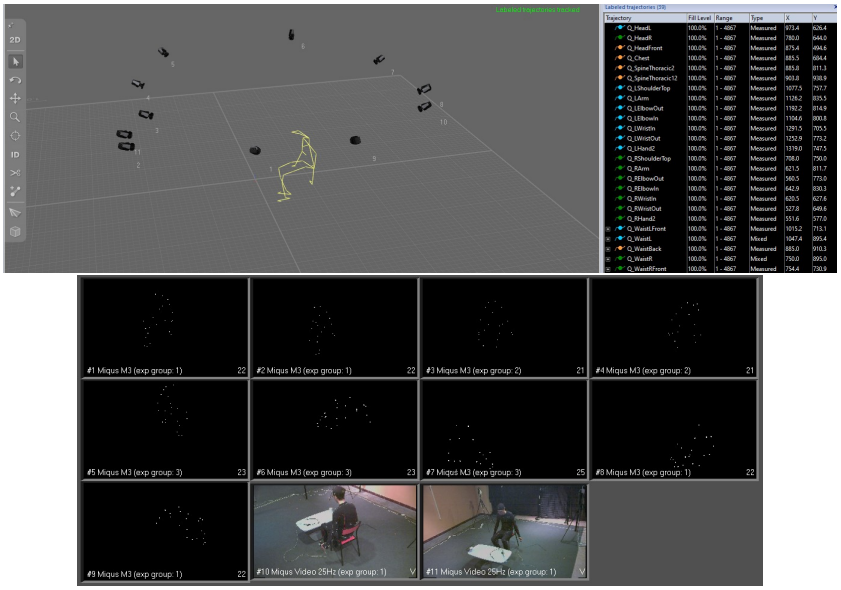


FIGURE 5.6 : Capture d'écran Qualisys Track Manager pendant l'acquisition de la séquence *Reach2*

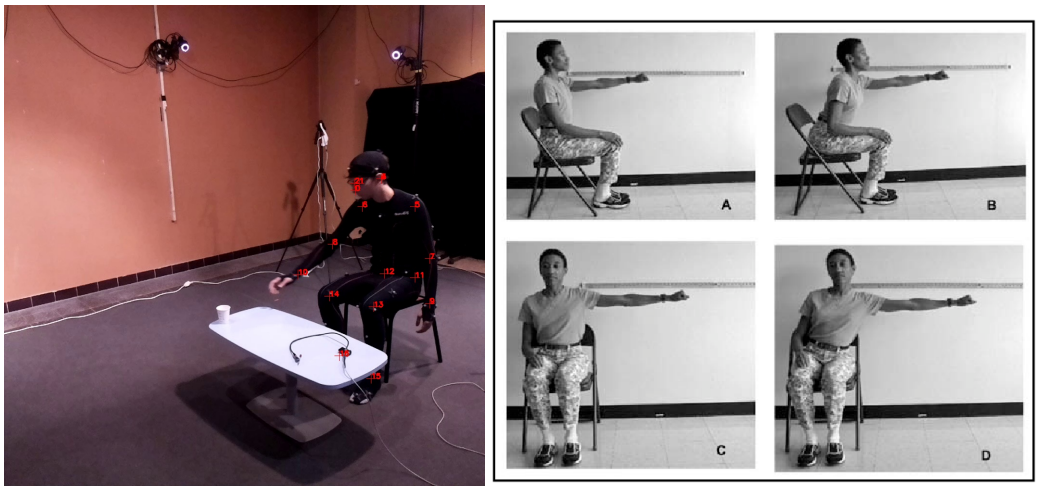


FIGURE 5.7 : Gauche : marqueurs suivis superposés avec les prédictions d'un modèle entraîné sur MS-COCO. Droite : la tâche de reaching dans Thompson and Medley (2007)

chaise. De plus, le test est effectué avec plusieurs objets dans la scène (table basse et chaise) comme lors des actions assises dans Human3.6m afin de tester la sensibilité aux occultations de l'ensemble du système.

5.5.2 Évaluer l'exactitude

Le modèle utilisé est celui sélectionné dans l'étape précédente comme le plus performant sur le jeu de donnée Human3.6M : *High resolution networks*. La version du modèle utilisée est pré-entraînée sur le jeu de donnée MS-COCO et utilise le groupement des points d'articulation par la méthode de Newell et al. (2017) (*associative embedding*). Les résultats obtenus en comparaison avec la capture de mouvement avec ce protocole sont reportés dans le tableau 5.5.

Articulation # images (regul.)	Reach1 817	Reach2 973	Reach3 1246	Reach1 817 (lms)	Reach2 973 (lms)	Reach3 1246 (lms)
ÉpauleG	69.93	71.44	73.83	70.73	71.16	66.13
ÉpauleD	54.95	54.31	52.90	55.64	54.47	55.48
CoudeG	27.71	26.01	36.24	25.29	23.26	24.85
CoudeD	51.63	50.31	50.91	52.71	48.86	50.48
PoignetG	115.63	142.77	85.58	79.96	77.86	39.32
PoignetD	69.12	61.28	95.24	51.58	50.41	81.41
TailleG	88.20	91.01	84.21	81.16	85.10	85.64
TailleD	98.39	65.08	67.67	103.96	72.50	79.76
GenouG	61.82	58.52	63.37	58.07	55.95	50.08
GenouD	92.96	82.12	130.24	91.37	79.67	99.70
PiedG	55.48	58.69	38.95	73.10	62.70	57.57
PiedD	49.02	58.55	54.32	55.44	66.60	62.96
Total :	**69.57**	**68.34**	**69.46**	**66.59**	**62.38**	**62.78**

TABLE 5.5 : Erreur du prototype avec le détecteur 2D *High resolution networks* en mm. MPJPE moyenne par articulation pour 3 séquence vidéos.

L'erreur moyenne obtenue est proche de 70 mm sur les 3 séquences capturées et 60 mm avec régularisation. Le prototype obtient une précision importante pour une tâche réputée complexe du fait de nombreuses occultations. Cette précision est obtenue sans régularisation ni *fine-tuning* du modèle sur les données utilisées pour l'évaluation. Il faut aussi noter que le placement des marqueurs ne correspond pas strictement aux annotations manuelles présentes dans le jeu de donnée MS-COCO.

Les articulations les mieux détectées sont les coudes, et les pieds, bien qu'ils soient aussi proches, voir plus que lors des actions assises dans Human3.6M. La triangulation

permet une précision augmentée pour ces articulations beaucoup moins bien repérées avec une méthode monoculaire. Il semblerait cependant que le poignet gauche (celui qui ne réalise pas l'action de saisir l'objet) et le genou droit soit moins bien détectés, peut être toujours à cause des occultations. Ce problème est réduit avec l'utilisation du *"least median of squares"* pendant la triangulation.

Archi. 2D	Method	MPJPE, mm
	Méthodes multivues	
-	Trumble et al. (2017)†	87.3
HgNet	Martinez et al. (2017) (Multiview)	57.0
HgNet	Pavlakos et al. (2017b)	56.9
-	Tome et al. (2018)	52.8
-	Huang et al. (2019)†	37.5
Res152	Qiu et al. (2019)	31.17
Res50	He et al. (2020) + tri	30.4
Res50	He et al. (2020) + RPSM	26.9
SmplNet	Bultmann and Behnke (2021)	29.8
HrNet	Notre approche + RANSAC	38.0
HrNet	Notre approche + LMS	32.1

TABLE 5.6 : Comparaison de l'exactitude de plusieurs méthodes multivues de l'état de l'art. Tous les résultats reportés sont obtenus sur Human3.6M (Ionescu et al., 2014) sans données d'entraînement supplémentaires et rapportés en MPJPE absolue (la plus basse est la meilleure). Lorsque les méthodes font l'usage d'une détection 2D, l'architecture du détecteur utilisé est précisé : HgNet : *"Hourglass Networks"*Newell et al. (2016), Res50/152 : *"Residual Networks"* (He et al., 2015), SmplNet : *"Simple Baselines 2D"* (Xiao et al., 2018) et HrNet : *"High Resolution Networks"* (Sun et al., 2019). † : méthodes utilisant des IMU.

Pour comparaison avec d'autres méthodes de l'état de l'art, le processus utilisé dans le prototype jusqu'à la triangulation a été testé sur Human3.6M. Les résultats obtenus sont reportés dans le tableau 5.6. Le protocole d'évaluation utilisé est le protocole d'origine proposée par les créateurs du jeu de donnée et est identique à celui des autres méthodes reportées : entraînement sur les séquences des sujets 1, 5, 6, 7 et 8 et évaluation sur les séquences des sujets 9 et 11. Pas d'alignement procruste à une articulation "racine". Les coordonnées obtenues sont exprimées dans le repère de la scène ou d'une caméra définie au préalable.

La version du prototype utilisant la triangulation combiné à l'algorithme LMS (*least median of square*) est compétitive avec les méthodes récentes de l'état de l'art. L'in-

fluence de l'architecture 2D utilisée semble importante pour l'ensemble des méthodes comparées. Les détecteurs récents réputés moins prompts à l'erreur permettent une meilleure reconstruction en trois dimensions des articulations. L'utilisation des architectures *High Resolution Networks* ou *Residual Networks* permet d'expliquer en partie la performance des meilleures méthodes multivues dont celle proposée.

Cependant, les méthodes qui surpassent de peu le résultat obtenu sur KeenMT sont diverses (Recursive Pictorial Structure Model, Epipolar Transformers...) mais utilisent parfois aussi le même type de triangulations, mais couplées avec RANSAC. La méthode de référence qui consiste à utiliser la DLT (Hartley and Sturm, 1997), semble pourtant produire de meilleures triangulations pour cette tâche avec LMS que RANSAC. De plus, LMS ne nécessite pas de paramètres définis manuellement pour un problème restreint à 4 vues (uniquement 11 combinaisons d'au moins 2 caméras).

5.5.3 Évaluer la robustesse

Bien que sans marqueurs et ne nécessitant pas de périphériques supplémentaires en dehors de 4 caméras, le système nécessite un certain nombre de contraintes pour fonctionner qu'il est important de préciser. L'analyse qualitative de la robustesse du système est abordée sous trois critères, du matériel nécessaire, de la calibration et de l'environnement d'acquisition.

Du point de vue du matériel utilisé, le système reste multivue et nécessite donc un placement de plusieurs caméras dans la scène afin de pouvoir capturer le sujet sous plusieurs angles. Il est important de pouvoir placer ces caméras de part et d'autres d'objets pouvant causer de fortes occultations comme des bureaux ou des tables. Enfin, il faut pouvoir orienter le champ de chaque caméra pour englober toute l'action du sujet si celui-ci se déplace.

Une autre contrainte importante du système est la nécessité d'une étape de calibration. Il est important de pouvoir calibrer le système à l'aide d'une mire comme précisé dans la sous-section 5.3.1. Cela réduit la flexibilité d'acquisition, car il faut effectuer à nouveau cette étape à chaque déplacement d'une seule des caméras du système.

Enfin, par rapport aux contraintes d'environnement, le système est fonctionnel avec plusieurs niveaux de luminosité et des volumes de capture de taille diverse (testés de $4\,\mathrm{m}^3$ à $10\,\mathrm{m}^3$).

5.5.4 Conclusion

Pour conclure, prototype KeenMT permet la reconstruction du mouvement humain à partir de vidéos capturé via quatre caméras calibrées et synchronisées. Le système ne nécessite pas de marqueurs et fonctionne sous tout type d'environnement avec lumière naturelle et occultation par d'une partie de la scène (chaise, bureau).

Le prototype a été validé sur des séquences vidéos capturées en parallèle à un système de capture du mouvement commercial. Ce test permet de valider la partie matérielle du prototype avec le choix des capteurs, la calibration, synchronisation et le protocole d'acquisition. Les résultats obtenus sont prometteurs et ouvrent la voie à des tests pour des applications en analyse du mouvement humain (posture, démarche, analyse des membres supérieurs et inférieurs). Une utilisation du système dans le cadre d'étude ou diagnostics médicaux (troubles musculosquelettiques, analyse de la posture au travail, etc.) est également envisageable.

Enfin, la méthode d'estimation de posture utilisée dans le prototype et, plus particulièrement, la méthode de triangulation des "points d'articulations" dans la scène, ont permis d'obtenir des résultats compétitifs par rapport à l'état de l'art des méthodes d'estimation de posture multivues. Cette comparaison sur les nombreuses séquences vidéos du jeu de données Human3.6M (Ionescu et al., 2014) permet de valider la partie logicielle du système et plus particulièrement l'estimation de posture 3D.

Chapitre 6

Conclusion

Contents

6.1 Contributions

Ce chapitre conclut cette thèse en synthétisant les contributions apportées à la tâche d'estimation de posture pour l'étude du mouvement humain. Particulièrement l'étude de l'exploitation des données multivues et la mise en place d'un prototype de système complet permettant l'acquisition jusqu'à l'extraction des points d'articulation. Enfin, ce chapitre énonce les perspectives de recherche ouvertes par ce système ainsi que des pistes d'améliorations possibles.

6.1.1 Fusion de données multivues

La première contribution de cette thèse est l'adaptation d'une méthode d'estimation de posture monoculaire Martinez et al. (2017) au cas multivue en le formulant comme un problème de fusion de donnée. Les images de chaque vue sont considérées comme les différentes sources de données pour extraire la posture. Ces données sont, dans l'état de l'art, souvent fusionnées à la fin d'un processus d'apprentissage qui les traite indépendamment (fusion tardive). Les expériences dans le cadre de ces contributions considèrent une fusion au début de ce processus (fusion aux données d'entrée), moins explorée dans le cadre de l'estimation de posture. Deux types de données sont utilisés pour la fusion puis l'apprentissage lors des tests. Dans un premier temps, les trajectoires 2D sont utilisées directement comme dans l'architecture d'origine. Un détecteur 2D de l'état de l'art permet d'extraire les coordonnées dans chaque vue. Dans un second temps, les silhouettes des personnes dans chaque vue sont aussi utilisées comme caractéristiques. Pour cela, un nouveau modèle de segmentation de personne est entraîné sur la base d'un algorithme de segmentation sémantique à plusieurs classes. Celui-ci peut également être utilisé indépendamment et fournit un compromis intéressant entre précision et fonctionnement en temps réel.

Ces deux architectures ont été testées et comparées à l'état de l'art sur le jeu de données Human3.6M avec les deux protocoles proposés par les auteurs de Ionescu et al. (2014) (coordonnées rectifiées ou non par rapport à un point de référence "racine" correspondant au marqueur central de la taille). L'expérience menée avec la fusion de coordonnée 2D a fourni des résultats peu concluants, surtout pour le protocole avec les coordonnées non alignées. La fusion de silhouettes n'a pas non plus permis une amélioration de la précision. Un dernier essai combinant les silhouettes et les coordonnées d'articulation produit des résultats légèrement meilleurs, mais n'améliorant toujours pas l'état de l'art. Ces résultats ont orienté le choix d'une méthode de fusion tardive pour le prototype décrit ci-après.

6.1.2 Prototype de système d'estimation de posture sans marqueurs

La contribution principale de cette thèse est le prototype KeenMT, un système de suivi tridimensionnel de la trajectoire du corps humain en mouvement. Cette contribution peut être résumée en deux parties. D'une part, le système d'acquisition qui permet la capture du mouvement d'une personne, de la mise en place des caméras et leur calibration jusqu'au traitement centralisé sur un ordinateur "poste de contrôle". D'autre part, la partie estimation de posture qui permet de traiter les données vidéos en multivue et d'en extraire les coordonnées 3D des articulations du sujet filmé.

Le système d'acquisition permet la capture depuis plusieurs angles de vue à l'aide de caméras branchées sur des ordinateurs monocartes qui assurent la synchronisation et la transmission des images à un serveur d'acquisition. Ce prototype introduit un système minimaliste et très peu couteux qui permet une capture du mouvement sans marqueurs aussi exacte que possible pour des composants faciles d'accès. Les caméras nécessitent une calibration pour permettre la mise en correspondance puis la triangulation de la posture.

Dans un second temps, une nouvelle méthode d'estimation de posture 3D adaptée au prototype pour permettre l'obtention de la précision optimale a été développée. Celle-ci est basée sur les meilleurs estimateurs de posture 2D existant et la triangulation par transformation linéaire directe (DLT). Le problème de la triangulation à partir de données multivues est la forte sensibilité aux valeurs aberrantes qui est souvent limitée par l'application d'un RANSAC pour éliminer les vues produisant des prédictions erronées. Dans nos essais, une variante basée sur le score de reprojection médian semble produire de meilleurs résultats.

Le prototype a été évalué dans sa totalité en condition réelle en le confrontant aux résultats obtenus à partir de capture de mouvement optique (avec marqueurs). Le mouvement capturé est semblable à une action étudiée en science du mouvement (reaching). La moyenne de l'erreur toutes articulations confondues par rapport à leur suivi avec marqueurs est de 63 mm sur les séquences vidéos capturées in situ et de 32.1 mm lorsque l'algorithme d'estimation de posture isolé est évalué sur tout le jeu de donnée Human3.6M. Ce résultat prometteur indique qu'une précision similaire aux méthodes de l'état de l'art pour l'estimation de posture humaine 3D à partir d'images multivue est atteignable en utilisant un tel système d'acquisition.

KeenMT démontre la faisabilité d'un système de capture du mouvement sans marqueurs utilisant des composants peu couteux tout en fournissant une précision acceptable pour l'étude de mouvements humains.

6.2 Limitations et perspectives

6.2.1 Limitation de la fusion multivues

Pour l'estimation de posture à partir de données multivues vidéo, les meilleures méthodes actuelles s'appuient sur la performance des méthodes d'estimation de posture 2D tout en exploitant la géométrie de la scène 3D. Elles utilisent souvent les méthodes de triangulation connues pour obtenir la profondeur des points d'articulations. Les expériences de fusion de donnée d'entrée tendent à confirmer que ce type d'approche est plus adapté au cas multivue. Cependant, le plafond d'exactitude qui aurait pu être atteint avec la méthode avec de meilleures détections 2D n'a pas été atteint lors de nos expériences. Les perspectives d'améliorations sont donc les suivantes :

- Coordonnées 3D en entrée
- Utilisation des séquences de posture
- Une architecture de convolution (avec en entrée des données plus complexes)
- Détecteur 2D plus performant

L'utilisation de coordonnée 3D ou d'autres types de représentations intermédiaires en 3D dimension (comme des *visual hull* qui pourrait être extrait des silhouettes) est la première possibilité. Le modèle affinerait donc la posture et permettrait de réduire les potentielles erreurs de triangulation dûes à des occultations. Une autre possibilité serait d'utiliser des séquences de postures et non plus des postures individuelles en adaptant l'architecture (par exemple avec *temporal convolutional networks* comme dans Pavllo et al. (2019)). De manière générale, des architectures plus complexes de réseaux de neurones profonds semblent plus adaptés que le modèle testé. Enfin, le détecteur 2D utilisé lors de la première étape pourrait être remplacé par un détecteur plus récent et plus précis.

Le problème principal de la fusion de donnée en entrée pour l'estimation de posture multivue est le biais introduit par les données d'apprentissage. En effet, les angles de vues utilisés ne changent souvent jamais, même dans les jeux de données à grande échelle. Cela introduit une sensibilité accrue pour des postures capturées à des endroits précis dans la scène et il est difficile de quantifier ce biais en dehors d'analyse par jeux de données croisés. Ce problème existe pour la plupart des méthodes basées sur l'apprentissage profond dans ce domaine, mais plus particulièrement pour la fusion en entrée, car les opérations de triangulation ou de fusion (et donc l'utilisation des paramètres extrinsèques des caméras) se font en amont de l'apprentissage.

Une des pistes explorée par He et al. (2020) et l'utilisation des paramètres des caméras pour améliorer les prédictions des estimateurs 2D. C'est une perspective peu explorée qui permettrait d'améliorer les résultats de toutes les méthodes à deux étapes qui se basent sur des coordonnées 2D.

6.2.2 Limitations du prototype

Les perspectives d'améliorations du prototype concernent d'une part le système d'acquisition et d'autre part la méthode d'estimation de posture :

Les améliorations matérielles qui pourraient être apportées peuvent permettre soit de simplifier ou réduire le prix du système ou encore d'en améliorer la robustesse ou le fonctionnement. Il est possible par exemple d'adapter le système avec des cartes plus simples pour réduire encore le coût global. Il est aussi possible de choisir des capteurs plus performants avec une meilleure résolution ou une plus forte fréquence d'acquisition pour, par exemple, capturer des mouvements plus rapides. La possibilité de mettre en place la communication sans-fil avec le serveur d'acquisition est par exemple envisageable pour permettre une installation plus facile. Se pose alors le problème du bon protocole de synchronisation et de communication entre les appareils. Jouer sur le nombre de caméras utilisées peut aussi permettre soit de réduire leur nombre et le coût du système (jusqu'à 2 caméras) mais risque de réduire la précision et la sensibilité aux occultations. À l'inverse, multiplier les capteurs et les angles de vue permet d'améliorer la reconstruction en augmentant les chances d'avoir plus d'images avec des détections non erronées ou occultées des articulations. Enfin, la possibilité de rendre le système complètement indépendant des capteurs utilisés serait une perspective qui améliorerait grandement l'accessibilité et les possibilités de déploiement du système.

L'étape de la calibration est également une problématique pour l'accessibilité du système, car c'est un processus long et fastidieux. Il existe cependant de plus en plus de propositions permettant la calibration manuelle à partir de quelques vidéos de mires de calibration (par exemple EasyMocap) mais il reste toujours difficile à appréhender. Une proposition serait d'intégrer un processus de calibration simplifié au système qui permettrait la modification rapide des angles de vues capturés.

Pour l'estimation de posture, il existe plusieurs limites du système actuel et aussi pour l'ensemble des méthodes de l'état de l'art entraîné sur les mêmes jeux de donnée. Ces méthodes d'apprentissage profond se basent sur des jeux de données qui contiennent un nombre important de postures différentes. Cependant, ils restent limités pour plusieurs raisons. Les mouvements sont souvent effectués par un nombre limité d'acteurs, ce qui introduit un biais pour des tailles, corpulence et apparence

dans les modèles entraînés. De plus, tous ces sujets filmés sont équipés de combinaison de capture de mouvement avec marqueurs, ce qui ajoute au biais d'apparence. Enfin, les actions que ces acteurs effectuent sont également contraintes par l'environnement contrôlé d'un volume d'acquisition en intérieur.

La méthode d'estimation de posture utilisée est basée sur un détecteur entraîné sur ce type de jeu de donnée, ce qui permet donc de relativiser l'exactitude de 30 mm obtenue sur le jeu de donnée contenant des images proches de son ensemble d'apprentissage. Cependant, le résultat de l'évaluation en laboratoire (63 mm) est aussi à considérer avec précaution à cause des variations que peuvent provoquer le placement des marqueurs lors de l'expérience.

L'autre défi principal de la méthode d'estimation de posture se situe lors de la triangulation des points d'articulation. Les erreurs de détections qui produisent des valeurs aberrantes dégradent fortement l'exactitude en 3D à la fin du processus. Malgré l'amélioration proposée plus haut, le problème persiste. Une possibilité à envisager serait de coupler cette suppression des valeurs aberrantes avec un processus d'amélioration de la précision 2D dans le cadre multivue comme He et al. (2020) ou à l'aide d'une mesure de confiance pondérée sur les différentes vues comme dans Iskakov et al. (2019).

6.3 Conclusion finale

Le prototype développé à l'issue de cette thèse permet l'estimation de la posture tridimensionnelle sans-marqueurs à partir de caméras à bas coût. La posture en 2D du sujet filmé est obtenue depuis des images à l'aide de détecteurs de posture 2D de l'état de l'art. Un système composé de multiples caméras calibrées limite le problème des occultations et permet la reconstruction de la posture en trois dimensions.

Ce système propose une solution simple et efficace au problème d'estimation de posture en vision par ordinateur avec une exactitude proche de l'état de l'art. De plus, il ouvre la voie à une capture de la posture facilitée pour tout type d'environnement dans de nombreux domaines (pour le sport et l'analyse de la performance, pour l'étude du mouvement humain ou pour la santé et la rééducation par exemple).

Grâce à l'intelligence artificielle, mais aussi aux connaissances actuelles en vision 3D, KeenMT permet l'estimation de la posture humaine pour un coût total du système bien inférieur aux méthodes avec marqueurs. Cela démontre la possibilité d'une démocratisation grandissante de l'accès à la capture du mouvement pour tout type d'usages.

Bibliographie

A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1) :44–58, January 2006. ISSN 1939-3539. doi : 10.1109/TPAMI.2006.21. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

JF Andersen, J Busck, and H Heiselberg. Submillimeter 3-d laser radar for space shuttle tile inspection. *Danisch Defense Research Establishment, Copenhagen, Denmark*, 2013.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, James Davis, and Jim Rodgers. SCAPE : Shape Completion and Animation of People. page 9, 2005.

Nicolas Aussel, Fabian Dubourvieux, and Yohan Petetin. Spatio-temporal convolutional neural networks for failure prediction. page 5, 2019.

So Young Baek, Mirel Ajdaroski, Payam Mirshams Shahshahani, Mélanie L. Beaulieu, Amanda O. Esquivel, and James A. Ashton-Miller. A Comparison of Inertial Measurement Unit and Motion Capture Measurements of Tibiofemoral Kinematics during Simulated Pivot Landings. *Sensors*, 22(12) :4433, January 2022. ISSN 1424-8220. doi : 10.3390/s22124433. URL `https://www.mdpi.com/1424-8220/22/12/4433`. Number : 12 Publisher : Multidisciplinary Digital Publishing Institute.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, April 2018. URL `http://arxiv.org/abs/1803.01271`. arXiv :1803.01271 [cs].

P. A. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure from motion. In Gerhard Goos, Juris Hartmanis, and Jan-Olof Eklundh, editors, *Computer Vision — ECCV '94*, volume 801, pages 85–96. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. ISBN 978-3-540-57957-1 978-3-540-48400-4. doi : 10.1007/BFb0028337. URL `http://link.springer.com/10.1007/BFb0028337`. Series Title : Lecture Notes in Computer Science.

Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi : 10.1109/CVPR.2014.216. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909612.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT : Real-Time Instance Segmentation. pages 9157–9166, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Bolya_YOLACT_Real-Time_Instance_Segmentation_ICCV_2019_paper.html.

Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6) :121, September 2021. ISSN 1432-1769. doi : 10.1007/s00138-021-01249-8. URL https://doi.org/10.1007/s00138-021-01249-8.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Simon Bultmann and Sven Behnke. Real-Time Multi-View 3D Human Pose Estimation using Semantic Feedback to Smart Edge Sensors. In *Robotics : Science and Systems XVII*, July 2021. doi : 10.15607/RSS.2021.XVII.040. URL http://arxiv.org/abs/2106.14729. arXiv :2106.14729 [cs].

M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, 2013.

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2272–2281, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi : 10.1109/ICCV.2019.00236. URL https://ieeexplore.ieee.org/document/9009459/.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, July 2017. doi : 10.1109/CVPR.2017.143. ISSN : 1063-6919.

Anargyros Chatzitofis, Georgios Albanis, Nikolaos Zioulis, and Spyridon Thermos. A Low-cost & Realtime Motion Capture System. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21421–21426, New Orleans,

LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi : 10.1109/CVPR52688.2022.02078. URL https://ieeexplore.ieee.org/document/9880370/.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. *arXiv :1711.07319 [cs]*, April 2018. URL http://arxiv.org/abs/1711.07319. arXiv : 1711.07319.

Yucheng Chen, Yingli Tian, and Mingyi He. Monocular Human Pose Estimation : A Survey of Deep Learning-based Methods. *Computer Vision and Image Understanding*, 192 :102897, March 2020. ISSN 10773142. doi : 10.1016/j.cviu.2019.102897. URL http://arxiv.org/abs/2006.01423. arXiv : 2006.01423.

Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T. Tan. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. pages 723–732, 2019. URL http://openaccess.thecvf.com/content_ICCV_2019/html/Cheng_Occlusion-Aware_Networks_for_3D_Human_Pose_Estimation_in_Video_ICCV_2019_paper.html.

Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training, 2020.

CMU. Graphic lab motion capture database. http://mocap.cs.cmu.edu/. URL http://mocap.cs.cmu.edu/.

Steffi L. Colyer, Murray Evans, Darren P. Cosker, and Aki I. T. Salo. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Medicine - Open*, 4, June 2018. ISSN 2199-1170. doi : 10.1186/s40798-018-0139-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986692/.

Stefano Corazza, L Mündermann, Ajit Chaudhari, T Demattio, C Cobelli, and Thomas Andriacchi. A Markerless Motion Capture System to Study Musculoskeletal Biomechanics : Visual Hull and Simulated Annealing Approach. *Annals of biomedical engineering*, 34 :1019–29, July 2006. doi : 10.1007/s10439-006-9122-8.

Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P. Andriacchi. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. *International Journal of Computer Vision*, 87(1) :156, September 2009. ISSN 1573-1405. doi : 10.1007/s11263-009-0284-3. URL https://doi.org/10.1007/s11263-009-0284-3.

Farzin Dadashi, Florent Crettenand, Grégoire P. Millet, and Kamiar Aminian. Front-Crawl Instantaneous Velocity Estimation Using a Wearable Inertial Measurement Unit. *Sensors*, 12(10) :12927–12939, October 2012. ISSN 1424-8220. doi : 10.3390/

s121012927. URL https://www.mdpi.com/1424-8220/12/10/12927. Number : 10 Publisher : Molecular Diversity Preservation International.

Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database, 2008. URL https://www.ri.cmu.edu/publications/guide-to-the-carnegie-mellon-university-multimodal-activity-cmu-mmac-da Library Catalog : www.ri.cmu.edu.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet : A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3D human pose estimation algorithms for markerless motion capture. *arXiv :2010.06449 [cs]*, July 2021. URL http://arxiv.org/abs/2010.06449. arXiv : 2010.06449.

Changxing Ding and Dacheng Tao. Robust Face Recognition via Multimodal Deep Face Representation. *IEEE Transactions on Multimedia*, 17(11) :2049–2058, November 2015. ISSN 1520-9210, 1941-0077. doi : 10.1109/TMM.2015.2477042. URL http://arxiv.org/abs/1509.00244. arXiv :1509.00244 [cs].

Pamela W. Duncan, Debra K. Weiner, Julie Chandler, and Stephanie Studenski. Functional Reach : A New Clinical Measure of Balance. *Journal of Gerontology*, 45(6) : M192–M197, 11 1990. ISSN 0022-1422. doi : 10.1093/geronj/45.6.M192.

Ethan Eade. Gauss-Newton / Levenberg-Marquardt Optimization. page 9, 2013.

Germain Faity, Denis Mottet, and Jérôme Froger. Validity and Reliability of Kinect v2 for Quantifying Upper Body Kinematics during Seated Reaching. *Sensors*, 22(7) :2735, January 2022. ISSN 1424-8220. doi : 10.3390/s22072735. URL https://www.mdpi.com/1424-8220/22/7/2735. Number : 7 Publisher : Multidisciplinary Digital Publishing Institute.

Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1) :55–79, January 2005. ISSN 1573-1405. doi : 10.1023/B:VISI.0000042934.15159.49. URL https://doi.org/10.1023/B:VISI.0000042934.15159.49.

Robert Fiker, Linda H. Kim, Leonardo A. Molina, Taylor Chomiak, and Patrick J. Whelan. Visual deep lab cut : A user-friendly approach to gait analysis. *Journal of Neuroscience Methods*, page 108775, 2020. ISSN 0165-0270. doi : https://doi.org/10.1016/j.jneumeth.2020.108775. URL http://www.sciencedirect.com/science/article/pii/S0165027020301989.

M.A. Fischler and R.A. Elschlager. The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, C-22(1) :67–92, January 1973. ISSN 0018-9340. doi : 10.1109/T-C.1973.223602. URL http://ieeexplore.ieee.org/document/1672195/.

Wolfgang Förstner and Eberhard Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, volume 6, pages 281–305. Interlaken, 1987.

Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi : A Large Multipurpose Motion and Video Dataset. *arXiv :2003.01888 [cs, eess]*, March 2020. URL http://arxiv.org/abs/2003.01888. arXiv : 2003.01888.

Paolo Giannini, Giulia Bassani, Carlo Alberto Avizzano, and Alessandro Filippeschi. Wearable Sensor Network for Biomechanical Overload Assessment in Manual Material Handling. *Sensors*, 20(14) :3877, January 2020. ISSN 1424-8220. doi : 10.3390/s20143877. URL https://www.mdpi.com/1424-8220/20/14/3877. Number : 14 Publisher : Multidisciplinary Digital Publishing Institute.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv :1406.2661 [cs, stat]*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv : 1406.2661.

K. Grauman, G. Shakhnarovich, and T. Darrell. A Bayesian approach to image-based visual hull reconstruction. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–187–I–194, Madison, WI, USA, 2003. IEEE Comput. Soc. ISBN 978-0-7695-1900-5. doi : 10.1109/CVPR.2003.1211353. URL http://ieeexplore.ieee.org/document/1211353/.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose : Dense Human Pose Estimation In The Wild. *arXiv :1802.00434 [cs]*, February 2018. URL http://arxiv.org/abs/1802.00434. arXiv : 1802.00434.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, West Nyack, 2004. ISBN 978-0-511-18618-9. URL http://qut.eblib.com.au/patron/FullRecord.aspx?p=256634. OCLC : 1044713766.

Richard I. Hartley and Peter Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2) :146–157, November 1997. ISSN 1077-3142. doi : 10.1006/cviu.1997.0547. URL https://www.sciencedirect.com/science/article/pii/S1077314297905476.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael Black. Resolving 3D Human Pose Ambiguities With 3D Scene Constraints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2282–2292, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi : 10.1109/ICCV.2019.00237. URL https://ieeexplore.ieee.org/document/9010321/.

D Hatfield and G Sheirman. Validation of an outdoor-based passive optoelectric motion capture system. In *ISBS-Conference Proceedings Archive*, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv :1512.03385 [cs]*, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv : 1512.03385.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN, January 2018. URL http://arxiv.org/abs/1703.06870. arXiv :1703.06870 [cs].

Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I. Yu. Epipolar Transformers. *arXiv :2005.04551 [cs]*, May 2020. URL http://arxiv.org/abs/2005.04551. arXiv : 2005.04551.

Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3D pose estimation. *arXiv :1711.08585 [cs]*, 11214 :69–86, 2018. doi : 10.1007/978-3-030-01249-6_5. URL http://arxiv.org/abs/1711.08585. arXiv : 1711.08585.

Chun Hao Huang, Benjamin Allain, Edmond Boyer, Jean-Sébastien Franco, Federico Tombari, Nassir Navab, and Slobodan Ilic. Tracking-by-Detection of 3D Human Shapes : from Surfaces to Volumes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8) :1994–2008, August 2018. doi : 10.1109/TPAMI.2017.2740308. URL https://hal.inria.fr/hal-01588272.

Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. DeepFuse : An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image. *arXiv :1912.04071 [cs]*, December 2019. URL http://arxiv.org/abs/1912.04071. arXiv : 1912.04071.

Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut : A deeper, stronger, and faster multi-person pose estimation model. *CoRR*, abs/1605.03170, 2016. URL http://arxiv.org/abs/1605.03170.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M : Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7) :

1325–1339, July 2014. ISSN 0162-8828, 2160-9292. doi : 10.1109/TPAMI.2013.248. URL http://ieeexplore.ieee.org/document/6682899/.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable Triangulation of Human Pose. *arXiv :1905.05754 [cs]*, May 2019. URL http://arxiv.org/abs/1905.05754. arXiv : 1905.05754.

W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5) :922–923, Sep 1976. doi : 10.1107/S0567739476001873. URL https://doi.org/10.1107/S0567739476001873.

Robert Kanko, Gerda Strutzenberger, Marcus Brown, Scott Selbie, and Kevin Deluzio. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. preprint, engrXiv, February 2020. URL https://osf.io/j4rbg.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE : Video Inference for Human Body Pose and Shape Estimation. *arXiv :1912.05656 [cs]*, December 2019a. URL http://arxiv.org/abs/1912.05656. arXiv : 1912.05656.

Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. *arXiv :1903.02330 [cs]*, April 2019b. URL http://arxiv.org/abs/1903.02330. arXiv : 1903.02330.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. *arXiv :1909.12828 [cs]*, September 2019. URL http://arxiv.org/abs/1909.12828. arXiv : 1909.12828.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6) :84–90, May 2017. ISSN 0001-0782, 1557-7317. doi : 10.1145/3065386. URL https://dl.acm.org/doi/10.1145/3065386.

Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations. *arXiv :1803.08244 [cs]*, March 2018. URL http://arxiv.org/abs/1803.08244. arXiv : 1803.08244.

Gregorij Kurillo, Evan Hemingway, Mu-Lin Cheng, and Louis Cheng. Evaluating the Accuracy of the Azure Kinect and Kinect v2. *Sensors (Basel, Switzerland)*, 22(7) :2469, March 2022. ISSN 1424-8220. doi : 10.3390/s22072469. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9002889/.

Joe Lallemand, Magdalena Szczot, and Slobodan Ilic. *Human Pose Estimation in Stereo Images*. July 2014. ISBN 978-3-319-08848-8. doi : 10.1007/978-3-319-08849-5_2.

A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2) :150–162, February 1994. ISSN 0162-8828, 2160-9292, 1939-3539. doi : 10.1109/34.273735.

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

S. H. Lee and Javier Civera. Robust Uncertainty-Aware Multiview Triangulation. *ArXiv*, August 2020. URL https://www.semanticscholar.org/paper/e12e59ffa7d72ae22c655065ecc1fe508127d398.

Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics*, 6(58) :eabe1315, September 2021. doi : 10.1126/scirobotics.abe1315. URL https://www.science.org/doi/10.1126/scirobotics.abe1315. Publisher : American Association for the Advancement of Science.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet : Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. *arXiv :1611.06612 [cs]*, November 2016. URL http://arxiv.org/abs/1611.06612. arXiv : 1611.06612.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO : Common Objects in Context, February 2015. URL http://arxiv.org/abs/1405.0312. arXiv :1405.0312 [cs].

Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention Mechanism Exploits Temporal Contexts : Real-Time 3D Human Pose Reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5063–5072, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi : 10.1109/CVPR42600.2020.00511. URL https://ieeexplore.ieee.org/document/9156272/.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD : Single Shot MultiBox Detector. *arXiv :1512.02325 [cs]*, 9905 :21–37, 2016. doi : 10.1007/978-3-319-46448-0_2. URL http://arxiv.org/abs/1512.02325. arXiv : 1512.02325.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL : a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6) :1–16, November 2015. ISSN 0730-0301, 1557-7368. doi : 10.1145/2816795.2818013. URL https://dl.acm.org/doi/10.1145/2816795.2818013.

Xiaojun Lu, Xu Duan, Xiuping Mao, Yuanyuan Li, and Xiangde Zhang. Feature Extraction and Fusion Using Deep Convolutional Neural Networks for Face Detection. *Mathematical Problems in Engineering*, 2017 :e1376726, January 2017. ISSN 1024-123X. doi : 10.1155/2017/1376726. URL https://www.hindawi.com/journals/mpe/2017/1376726/. Publisher : Hindawi.

Fabricio Anicio de Magalhaes, Giuseppe Vannozzi, Giorgio Gatta, and Silvia Fantozzi. Wearable inertial sensors in swimming motion analysis : a systematic review. *Journal of Sports Sciences*, 33(7) :732–745, April 2015. ISSN 0264-0414. doi : 10.1080/02640414.2014.962574. URL https://doi.org/10.1080/02640414.2014.962574. Publisher : Routledge _eprint : https ://doi.org/10.1080/02640414.2014.962574.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass : Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL https://amass.is.tue.mpg.de.

Académie des sciences (France) Marey. Comptes rendus hebdomadaires des séances de l'Académie des sciences / publiés... par MM. les secrétaires perpétuels, January 1882. URL https://gallica.bnf.fr/ark:/12148/bpt6k3050z.

Etienne-Jules Marey. La chronophotographie : nouvelle méthode pour analyser le mouvement dans les sciences pures et naturelles. *Revue générale des sciences pures et appliquées*, 2, 1891.

E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial human spaces : How well do humans perceive a 3d articulated pose? In *2013 IEEE International Conference on Computer Vision*, pages 1289–1296, 2013.

Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces : A computational study on the human perception of 3d articulated poses. *International Journal of Computer Vision*, 119(2) :194–215, Apr 2016. doi : 10.1007/s11263-016-0888-3. URL http://dx.doi.org/10.1007/s11263-016-0888-3.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. *arXiv :1705.03098 [cs]*, August 2017. URL http://arxiv.org/abs/1705.03098. arXiv : 1705.03098.

Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut : markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9) :1281–1289, September 2018. ISSN 1546-1726. doi : 10.1038/s41593-018-0209-y. URL https://www.nature.com/articles/s41593-018-0209-y. Number : 9 Publisher : Nature Publishing Group.

Mackenzie W. Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *arXiv :1909.13868 [cs, q-bio]*, October 2019. URL http://arxiv.org/abs/1909.13868. arXiv : 1909.13868.

Ruth E. Mayagoitia, Anand V. Nene, and Peter H. Veltink. Accelerometer and rate gyroscope measurement of kinematics : an inexpensive alternative to optical motion analysis systems. *Journal of Biomechanics*, 35(4) :537–542, April 2002. ISSN 0021-9290. doi : 10.1016/s0021-9290(01)00231-7.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. [1611.09813] Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision, 2016. URL https://arxiv.org/abs/1611.09813.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect : real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics*, 36(4) :1–14, July 2017. ISSN 07300301. doi : 10.1145/3072959.3073596. URL http://dl.acm.org/citation.cfm?doid=3072959.3073596.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. *arXiv :1712.03453 [cs]*, August 2018. URL http://arxiv.org/abs/1712.03453. arXiv : 1712.03453.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect : Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Transactions on Graphics*, 39(4), July 2020. ISSN 0730-0301, 1557-7368. doi : 10.1145/3386569.3392410. URL http://arxiv.org/abs/1907.00837. arXiv : 1907.00837.

Thomas B. Moeslund and Erik Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3) :231–268, March 2001. ISSN 1077-3142. doi : 10.1006/cviu.2000.0897. URL http://www.sciencedirect.com/science/article/pii/S107731420090897X.

Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2V-PoseNet : Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map, August 2018. URL http://arxiv.org/abs/1711.07399. arXiv :1711.07399 [cs].

Greg Mori, Xiaofeng Ren, Alexei A Efros, and Jitendra Malik. Recovering Human Body Configurations : Combining Segmentation and Recognition. page 8, 2004.

Matteo Moro, Giorgia Marchesi, Francesca Odone, and Maura Casadio. Markerless gait analysis in stroke survivors based on computer vision and deep learning : A pilot study. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 2097–2104, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368667. doi : 10.1145/3341105.3373963. URL https://doi.org/10.1145/3341105.3373963.

Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. preprint, Neuroscience, November 2018. URL http://biorxiv.org/lookup/doi/10.1101/476531.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. *arXiv :1603.06937 [cs]*, March 2016. URL http://arxiv.org/abs/1603.06937. arXiv : 1603.06937.

Alejandro Newell, Zhiao Huang, and Jia Deng. Associative Embedding : End-to-End Learning for Joint Detection and Grouping, June 2017. URL http://arxiv.org/abs/1611.05424. arXiv :1611.05424 [cs].

Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, Bhubaneswar, India, December 2008. IEEE. doi : 10.1109/ICVGIP.2008.47. URL http://ieeexplore.ieee.org/document/4756141/.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural Body Fitting : Unifying Deep Learning and Model-Based Human Pose and Shape Estimation. *arXiv :1808.05942 [cs]*, August 2018. URL http://arxiv.org/abs/1808.05942. arXiv : 1808.05942.

Joseph O'Rourke and Norman Badler. Model-Based Image Analysis of Human Motion Using Constraint Propagation. *PAMI*, 2 :522–536, November 1980. doi : 10.1109/TPAMI.1980.6447699.

Joseph O'Rourke and Norman I Badler. Image Analysis of Human Motion Using Constraint Propagation. page 61, 1979.

George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards Accurate Multi-person Pose Estimation in the Wild, April 2017. URL http://arxiv.org/abs/1701.01779. arXiv :1701.01779 [cs].

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. *arXiv :1611.07828 [cs]*, July 2017a. URL http://arxiv.org/abs/1611.07828. arXiv : 1611.07828.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting Multiple Views for Marker-less 3D Human Pose Annotations, April 2017b. URL http://arxiv.org/abs/1704.04793. arXiv :1704.04793 [cs].

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi : 10.1109/CVPR.2019.00794. URL https://ieeexplore.ieee.org/document/8954163/.

Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross View Fusion for 3D Human Pose Estimation. *arXiv :1909.01203 [cs]*, September 2019. URL http://arxiv.org/abs/1909.01203. arXiv : 1909.01203.

Srikumar Ramalingam, Suresh K. Lodha, and Peter Sturm. A generic structure-from-motion framework. *Computer Vision and Image Understanding*, 103(3) :218–228, September 2006. ISSN 10773142. doi : 10.1016/j.cviu.2006.06.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1077314206000695.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once : Unified, Real-Time Object Detection. *arXiv :1506.02640 [cs]*, May 2016. URL http://arxiv.org/abs/1506.02640. arXiv : 1506.02640.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv :1505.04597 [cs]*, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv : 1505.04597.

Marco Seeland, Michael Rzanny, Nedal Alaqraa, Jana Wäldchen, and Patrick Mäder. Plant species classification using flower images—A comparative study of local feature representations. *PLOS ONE*, 12 :e0170629, February 2017. doi : 10.1371/journal.pone.0170629.

Swathi Sheshadri, Benjamin Dann, Timo Hueser, and Hansjoerg Scherberger. 3D reconstruction toolbox for behavior tracked with multiple cameras. *Journal of Open Source Software*, 5(45) :1849, January 2020. ISSN 2475-9066. doi : 10.21105/joss.01849. URL https://joss.theoj.org/papers/10.21105/joss.01849.

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. page 8.

Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva : Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision*, 87(1-2) :4–27, March 2010. ISSN 0920-5691, 1573-1405. doi : 10.1007/s11263-009-0273-6. URL http://link.springer.com/10.1007/s11263-009-0273-6.

Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos, November 2014. URL http://arxiv.org/abs/1406.2199. arXiv :1406.2199 [cs].

Rim Slama, Oussama Ben-Ammar, Houda Tlahig, Ilhem Slama, and Pierre Slangen. Human-centred assembly and disassembly systems : a survey on technologies, ergonomic, productivity and optimisation. *IFAC-PapersOnLine*, 55(10) :1722–1727, January 2022. ISSN 2405-8963. doi : 10.1016/j.ifacol.2022.09.646. URL https://www.sciencedirect.com/science/article/pii/S2405896322019632.

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

Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional Human Pose Regression. *arXiv :1704.00159 [cs]*, August 2017. URL http://arxiv.org/abs/1704.00159. arXiv : 1704.00159.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral Human Pose Regression. *arXiv :1711.08229 [cs]*, September 2018. URL http://arxiv.org/abs/1711.08229. arXiv : 1711.08229.

Mary Thompson and Ann Medley. Forward and lateral sitting functional reach in younger, middle-aged, and older adults. *Journal of Geriatric Physical Therapy*, 30 : 43–48, 2007.

Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking Pose in 3D : Multi-stage Refinement and Recovery for Markerless Motion Capture. *arXiv :1808.01525 [cs]*, August 2018. URL http://arxiv.org/abs/1808.01525. arXiv : 1808.01525.

Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *arXiv :1406.2984 [cs]*, June 2014. URL http://arxiv.org/abs/1406.2984. arXiv : 1406.2984.

Alexander Toshev and Christian Szegedy. DeepPose : Human Pose Estimation via Deep Neural Networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014. doi : 10.1109/CVPR.2014.214. URL http://arxiv.org/abs/1312.4659. arXiv : 1312.4659.

Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture : 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *Procedings of the British Machine Vision Conference 2017*, page 14, London, UK, 2017. British Machine Vision Association. ISBN 978-1-901725-60-5. doi : 10.5244/C.31.14. URL http://www.bmva.org/bmvc/2017/papers/paper014/index.html.

S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4) : 376–380, 1991. doi : 10.1109/34.88573.

Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *Transactions on Pattern Analysis and Machine Intelligence*, 38(8) : 1533–1547, January 2016. doi : 10.1109/TPAMI.2016.2522398. URL http://dx.doi.org/10.1109/TPAMI.2016.2522398.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11214, pages 614–631. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01248-9 978-3-030-01249-6. doi : 10.1007/978-3-030-01249-6_37. URL http://link.springer.com/10.1007/978-3-030-01249-6_37. Series Title : Lecture Notes in Computer Science.

Bastian Wandt and Bodo Rosenhahn. RepNet : Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. *arXiv :1902.09868 [cs]*, March 2019. URL http://arxiv.org/abs/1902.09868. arXiv : 1902.09868.

Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion Guided 3D Pose Estimation from Videos. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12358, pages 764–780. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58600-3 978-3-030-58601-0. doi : 10.1007/978-3-030-58601-0_45. URL http://link.springer.

com/10.1007/978-3-030-58601-0_45. Series Title : Lecture Notes in Computer Science.

Cmglee Wikimedia Commons. Comparison of the lenet and alexnet convolution, pooling, and dense layers. https://commons.wikimedia.org/wiki/File:Comparison_image_neural_networks.svg, 2021. URL https://commons.wikimedia.org/wiki/File:Comparison_image_neural_networks.svg.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking, 2018.

Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC : Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. *arXiv :1910.00116 [cs, eess]*, October 2019. URL http://arxiv.org/abs/1910.00116. arXiv : 1910.00116.

Sung-Pyo Yang, Yeong-Hyeon Seo, Jae-Beom Kim, Henry Hyunwoo Kim, and Ki-Hun Jeong. Optical MEMS devices for compact 3D surface imaging cameras. *Micro and Nano Systems Letters*, 7, December 2019. doi : 10.1186/s40486-019-0087-4.

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion : Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 910–919, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi : 10.1109/ICCV.2017.104. URL http://ieeexplore.ieee.org/document/8237366/.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. DoubleFusion : Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor. *arXiv :1804.06023 [cs]*, April 2018. URL http://arxiv.org/abs/1804.06023. arXiv : 1804.06023.

Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Han Xi, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu. Pose2Seg : Detection Free Human Instance Segmentation, April 2019. URL http://arxiv.org/abs/1803.10683. arXiv :1803.10683 [cs].

Yilun Zhang and Hyun Soo Park. Multiview Supervision By Registration. pages 420–428, 2020. URL http://openaccess.thecvf.com/content_WACV_2020/html/Zhang_Multiview_Supervision_By_Registration_WACV_2020_paper.html.

Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11) :1330–1334, November 2000. ISSN 01628828. doi : 10.1109/34.888718. URL http://ieeexplore.ieee.org/document/888718/.

Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation : A geometric approach, 2020.

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D Human Pose Estimation in the Wild : a Weakly-supervised Approach. *arXiv :1704.02447 [cs]*, July 2017. URL `http://arxiv.org/abs/1704.02447`. arXiv : 1704.02447.