



Contribution à la capture du mouvement humain par stéréovision et machine learning pour l'analyse de la marche

Saman Vafadar

► To cite this version:

Saman Vafadar. Contribution à la capture du mouvement humain par stéréovision et machine learning pour l'analyse de la marche. Biomécanique [physics.med-ph]. HESAM Université, 2020. Français. NNT : 2020HESAE069 . tel-03390546

HAL Id: tel-03390546

<https://pastel.hal.science/tel-03390546>

Submitted on 21 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE SCIENCES DES MÉTIERS DE L'INGÉNIEUR
Institut de Biomécanique Humaine Georges Charpak – Campus de Paris

THÈSE

présentée par : **Saman VAFADAR**

soutenue le : 16 Décembre 2020

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée à : **École Nationale Supérieure d'Arts et Métiers**

Spécialité : Biomécanique

Contribution à la capture du mouvement humain par stéréovision et machine learning pour l'analyse de la marche

THÈSE dirigée par :
Mme SKALLI Wafa

et co-encadrée par :
M. GAJNY Laurent

Jury

Mme Agnès LINGLART, Professeur des Universités, Université Paris-Saclay
M. Floren COLLOUD, Maître de conférences HDR, Université de Poitiers
M. Joao Manuel R. S. TAVARES, Professeur associé HDR, Université de Porto
M. Olivier GIBARU, Professeur des Universités, ENSAM-Lille
M. Ayman ASSI, Professeur associé, Université Saint-Joseph de Beyrouth
M. Laurent GAJNY, Maître de conférences, ENSAM-Paris
Mme Wafa SKALLI, Professeur des Universités, ENSAM-Paris

Présidente
Rapporteur
Rapporteur
Examinateur
Examinateur
Examinateur
Examinatrice

**T
H
È
S
E**

Acknowledgement

First and foremost, I would like to express my profound gratitude to my supervisors Prof. Wafa Skalli (Ph.D.) and Dr. Laurent Gajny (Ph.D.), for their candid guidance, constructive propositions, and incredible motivation in nurturing the research. It has been a blessing for me to spend many opportune moments under their direction since the inception of the thesis. In particular, I genuinely appreciate the full support, willingness to help, continuous guidance, inspiring ideas of Dr. Gajny during these three challenging years. I could not finish this Ph.D. without his invaluable help. Apart from professional activities, I learned many personal traits during our discourse facet-to-face or otherwise.

I would like to express my thanks to all the members of the jury. I want to thank Prof. Floren Colloud (Ph.D.) and Prof. Joao Manuel R. S. Tavares to agree to review this manuscript. I want to thank Prof. Agn  s Linglart (MD., Ph.D.), Prof. Olivier Gibaru (Ph.D.), Prof. Ayman Assi (Ph.D.) for having agreed to participate in the jury as examiners.

I would like to offer my sincere thanks to people in the lab without whom achieving the objectives of this Ph.D. would not be possible. Thank you, Prof. Helene Pillet (Ph.D.), Dr. Xavier Bonnet (Ph.D.) for providing crucial resources and information. Thank you, Mr. Sylvain Persohn (Eng.), for your assistance with the technical issues that I encountered. Thank you, Ms. Laura Valdes (Eng.), for your help with the motion capture and medical imaging systems. Thank you, Dr. Marc Khalif   (MD.), Dr. Amine Hamza (MD.) for your cooperation with in-vivo data acquisitions.

Special thanks and gratitude to Ms. Aurore Bonnet-Lebrun (MSc.) Her help and cooperation were invaluable to me. I genuinely appreciate her assistance and cooperation with the sessions of in-vivo data acquisition. Also, I would like to express my gratitude to Mr. Matthieu Bo  ss   (MSc.) who his attitude was positive and motivated, and helped us with developing new scientific solutions for an important challenge of this Ph.D. Ms. Marine Souq, please accept my deepest thanks for helping me with any administrative challenge that I faced during these three years. Mr. Mohamed Marhoum, I want you to know how much I value the conversations we occasionally had. I am thankful to you for your supportive thoughts.

I am grateful to Institut de Biom  canique Humaine Georges Charpak, BiomecAM chair program and its investors Fondation ParisTech, Soci  t   G  n  rale, COVEA for providing financial assistance. Finally, I would like to express my sincere gratitude to all who directly or indirectly helped me complete my thesis.

I would like to sincerely thank and express my respect to Ms. Alice Siegel (Ph.D.). The best months of my Ph.D. were when we shared the same office. I genuinely enjoyed our discussions on politics, France, Iran, and culture. Thanks to Mr. Bhrigu Lahkar Kumar (Ph.D.), who has been a kind and gentle colleague. Special thanks to Mr. Christophe Muth-Seng (Ph.D. very soon!), who has always been a welcoming and generous colleague.

Abstract

Abstract in English Gait analysis is the measurement of and ability evaluation of walking that can be used for fall risk assessment or as a diagnostic and prognostic tool for clinical applications. However, despite the clinical value, several difficulties attributed to the current established gold standard instrumentation, marker-based motion capture systems, limit the large-scale use in clinical applications. The current marker-based systems are costly and require a controlled laboratory environment. The test procedure is also time-consuming. Eliminating the markers would drastically shorten the patient preparatory time and would be more efficient. The objective of this study is to design a marker-less motion capture system for clinical applications. Recent advancements in computer vision and especially in convolutional neural networks, have provided the potential to pursue this objective. The designed system consists of four RGB cameras and can estimate the position of joint centers through a deep learning approach. For that purpose, a novel specific dataset has been collected including asymptomatic and pathologic subjects. To evaluate the validity of the developed system, its performance is assessed against a marker-based motion capture system in terms of joint position errors and clinically relevant gait parameters. The results demonstrate the high potential of the designed system for clinical applications.

Keywords – marker-less motion capture, gait analysis, pose estimation, deep learning, machine learning, stereovision.

Résumé en français L'analyse de la marche est la mesure et l'évaluation de la capacité de marche qui peut être utilisée pour l'évaluation des risques de chute ou comme outil de diagnostic et de pronostic pour des applications cliniques. Toutefois, malgré la valeur clinique, plusieurs difficultés attribuées à l'instrumentation de référence actuelle, les systèmes de capture du mouvement basés sur des marqueurs, limitent l'utilisation à grande échelle dans les applications cliniques. Les systèmes actuels sont coûteux et nécessitent un environnement de laboratoire contrôlé. La procédure de test est également longue. L'élimination des marqueurs réduirait considérablement le temps de préparation du patient et serait plus efficace. L'objectif de cette étude est de concevoir un système de capture de mouvement sans marqueur pour les applications cliniques. Les récents progrès réalisés dans le domaine de la vision par ordinateur et en particulier dans celui des réseaux neuronaux convolutifs, ont permis de poursuivre cet objectif. Le système conçu se compose de quatre caméras RGB et peut estimer la position des centres communs grâce à une approche d'apprentissage approfondie. À cette fin, un nouvel ensemble de données spécifiques a été collecté, incluant des sujets asymptomatiques et pathologiques. Pour évaluer la validité du système développé, ses performances sont évaluées par rapport à un système de capture de mouvement basé sur des marqueurs en termes d'erreurs de position des articulations et de paramètres de marche cliniquement pertinents. Les résultats démontrent le potentiel élevé du système conçu pour des applications cliniques.

Mots clés – capture de mouvement sans marqueur, analyse de la marche, estimation de la pose, apprentissage profond, machine learning, stéréovision.

Contents

Acknowledgement	i
Résumé Long en Français	v
R.1 Analyse quantitative de la marche	vii
R.2 Estimation de la pose humaine	xvi
R.3 Base de données ENSAM de poses humaines	xxii
R.4 Analyse quantitative de la marche sans marqueurs : validité et fiabilité	xxix
General Introduction	1
1 Gait Analysis	3
1.1 Gait analysis	4
1.1.1 Kinematic gait parameters	4
1.1.2 Spatiotemporal gait parameters	8
1.1.3 Role of gait analysis	9
1.1.4 Reliability of kinematic gait parameters	10
1.2 Limitations of gait analysis laboratories	13
1.3 Marker-less and IMU-based Motion capture systems	14
1.3.1 IMU-based motion capture systems	15
1.3.2 Marker-less motion capture systems	16
Depth cameras	16
RGB cameras	18
1.4 Conclusion	23
2 Human Pose Estimation	25
2.1 Human pose datasets	25
MPII dataset	26
CMU dataset	26
HumanEva dataset	27
Human3.6M dataset	27
2.2 Evaluation metrics	28
2.3 Evolution of human pose estimation	29
2.3.1 Features and representation	29
2.3.2 Representation learning	32
2.3.3 Convolutional neural networks	32
Pooling	33
2.4 Human body model	34
Kinematic model	35
Planar model	35
Volumetric model	35
2.5 Human pose estimation: State-of-the-arts	36
2.5.1 2D human pose estimation methods	36
2.5.2 3D human pose estimation methods	40

2.6	Conclusion	43
3	ENSAM pose dataset	45
3.1	Materials and methods	46
3.1.1	Reference marker-based motion capture system	46
3.1.2	Marker-less motion capture system	46
3.1.3	Experimental setup	47
3.1.4	Calibration	48
	Reference marker-based motion capture system	48
	Marker-less motion capture system	48
3.1.5	Synchronization	50
3.1.6	Data collection	51
3.1.7	Annotation data	52
	Joint centers	52
	Bounding boxes	53
3.1.8	Walking trials	55
3.2	Results	55
3.2.1	Calibration	55
3.2.2	Synchronization	58
3.2.3	Population	58
3.2.4	Data processing	59
3.3	Discussion	59
3.3.1	Experimental setup	60
3.3.2	Calibration	61
3.3.3	Synchronization	62
3.3.4	Data processing	62
3.4	Conclusion	63
4	Marker-less motion capture system: validity and reliability	65
4.1	Materials and methods	65
4.1.1	Human pose estimation	65
4.1.2	Training and test of the human pose estimation method	67
4.1.3	Detection of gait events	68
4.1.4	Spatiotemporal gait parameters	68
4.1.5	Kinematic gait parameters	69
4.1.6	Reliability of kinematic gait parameters	70
4.2	Results	71
4.3	Discussion	76
4.3.1	Human pose estimation	76
4.3.2	Detection of gait events	78
4.3.3	Spatiotemporal gait parameters	79
4.3.4	Kinematic gait parameters	80
4.3.5	Limitations	81
4.4	Conclusion	81
	General Conclusion and Perspectives	83
A	Pinhole camera model	85
B	Marker-set	87
	Bibliography	89

Résumé Long en Français

Introduction générale

La marche joue un rôle essentiel dans la qualité de vie et certaines pathologies ou simplement le vieillissement peuvent l'altérer. L'analyse quantitative de la marche (AQM) est la mesure et l'évaluation de la capacité à marcher (Davis et al., 1991). Parmi les paramètres mesurés lors de cette analyse, on retrouve par exemple des paramètres cinématiques et spatiotemporels.

Diverses études soutiennent l'utilisation de l'AQM pour l'évaluation des risques de chute et la prédiction des futures chutes dans la population adulte âgée. Parmi les personnes âgées de plus de 65 ans, plus de la moitié des hospitalisations sont liées à des chutes (WHO, 2008). Les chutes peuvent provoquer une fracture de la hanche, ce qui a des conséquences négatives sur la qualité de vie et représente un coût non-négligeable (Williamson et al., 2017). Les paramètres spatiotemporels de la marche, par exemple, la longueur de la foulée, sont d'une grande valeur pour l'évaluation des risques de chute. La longueur de foulée mesure la distance parcourue entre les contacts successifs d'un pied avec le sol. Par exemple, une augmentation de la variabilité de la longueur de foulée observée sur une période d'un an augmente la probabilité de chute (Hausdorff, Rios, and Edelberg, 2001).

De plus, l'AQM comme outil de diagnostic et de pronostic pour plusieurs conditions cliniques telles que la paralysie cérébrale et les lésions cérébrales acquises, est soutenue par la littérature (Benedetti et al., 2017). (Wren et al., 2011) a démontré que l'utilisation de l'AQM en plus des examens habituels a entraîné des changements dans les stratégies de traitement pour 52-89% des patients et 41-51% des procédures. En particulier, 37-39% des opérations prévues n'ont pas été réalisées, et 28-40% des opérations réalisées n'étaient pas prévues avant l'AQM.

Cependant, l'AQM présente bien sûr des inconvénients. Les limitations des laboratoires d'analyse de la marche sont principalement attribuées aux instruments actuellement considérés comme des références. Ceux-ci sont des systèmes de capture du mouvement basés sur des marqueurs. Ces systèmes peuvent retrouver avec précision la position d'un ensemble de marqueurs placés sur la surface de la peau du sujet et qui aident à calculer les paramètres de la marche. Toutefois, le coût de l'équipement, le salaire du personnel, la nécessité d'un environnement de laboratoire contrôlé et la procédure de test qui prend beaucoup de temps sont parmi les facteurs qui limitent l'utilisation généralisée de l'AQM en routine clinique. Pour ces raisons, plusieurs systèmes alternatifs de capture du mouvement ont été proposés.

Les systèmes de capture du mouvement basés sur des centrales inertielles (IMU – Inertial Measurement Unit) ou sans marqueurs figurent parmi les alternatives proposées. Bien que les systèmes de capture du mouvement basés sur les IMU ne nécessitent pas d'environnement de laboratoire, certaines difficultés en termes de précision de ces systèmes limitent leur utilisation dans les applications cliniques. Les systèmes de capture du mouvement sans marqueurs, basés sur des caméras de profondeur ou des caméras RGB (Red - Green - Blue), estiment les poses humaines en utilisant diverses techniques de vision par ordinateur (appelées méthodes d'estimation des poses humaines). Néanmoins, aucune étude publiée ne soutient la validité des

systèmes de capture du mouvement sans marqueurs pour les applications cliniques. L'objectif de cette étude est de développer un système de capture du mouvement sans marqueurs qui pourrait aider à généraliser l'utilisation de l'analyse quantitative de la marche dans les cliniques.

R.1 Analyse quantitative de la marche

Un cycle de marche commence par le contact du talon et se termine par le contact successif du même pied. L'action de lever des orteils divise le cycle de marche en phases d'appui et phase oscillante. La phase d'appui se réfère à la période entière pendant laquelle le pied reste en contact avec le sol. D'autre part, la phase oscillante fait référence à la période pendant laquelle le pied est hors du sol. La division classique d'un cycle de marche est illustrée dans la figure 1.1.

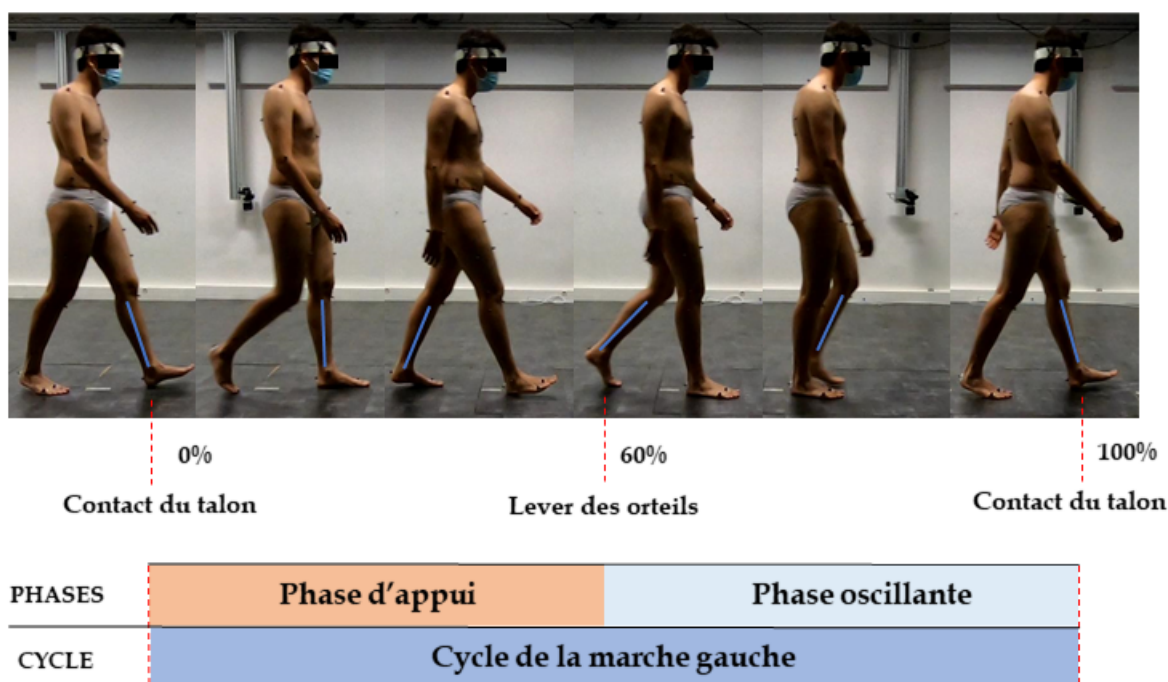


FIGURE 1: Divisions du cycle de marche gauche

Parmi les paramètres mesurés lors d'une AQM, on retrouve par exemple des paramètres cinématiques et spatiotemporels. **Les paramètres de marche cinématiques** (cinématique des articulations) représentent le mouvement relatif entre des os adjacents (Cappozzo et al., 2005). **Les paramètres spatio-temporels de la marche** sont des descripteurs spatio-temporels de la marche. Les systèmes de référence actuellement établis pour l'analyse clinique de la marche sont les suivants : **systèmes de capture du mouvement basés sur des marqueurs** et **plate-formes de force**. Ces systèmes peuvent aider à retrouver les paramètres de la marche qui nous intéressent dans ce travail de thèse. Dans les paragraphes suivants, nous expliquons comment ces paramètres de la marche sont mesurés à partir du système de référence.

Paramètres cinématiques de la marche

Un ensemble de marqueurs est placé sur la surface de la peau des segments du corps (par exemple, le tibia, le fémur et le bassin) de manière à estimer leur orientation. La position tridimensionnelle des marqueurs est récupérée par le système de capture du mouvement basé sur les marqueurs. Pour chaque segment du corps, deux types de systèmes de coordonnées sont définis, les repères **techniques** et **anatomiques**. Le repère technique, formé par trois marqueurs non alignés, vise à enregistrer l'orientation du segment corporel, dans l'espace tridimensionnel, à chaque instant. L'emplacement du repère technique est facultatif par rapport aux segments osseux sous-jacents et ne peut être répété par la suite. D'autre part, les repères anatomiques

sont définis par les points anatomiques, et répondent ainsi à l'exigence de répétabilité inter- et intra-sujet. Il existe différentes techniques pour estimer la position des points anatomiques et les enregistrer dans les repères techniques. Par exemple, les points anatomiques peuvent être identifiés par palpation, et leur position peut être déterminée en plaçant des marqueurs sur eux.

Pour la description de la cinématique des articulations, il faut considérer les repères anatomiques de deux segments osseux adjacents, le proximal (p) et le distal (d). Si ${}^gR_p = [p_X, p_Y, p_Z]$ et ${}^gR_d = [d_X, d_Y, d_Z]$ représentent les matrices d'orientation des segments osseux proximal et distal par rapport au système de coordonnées global (g), la matrice d'orientation de l'articulation peut être obtenue en utilisant l'équation 1.1 (Cappozzo et al., 2005).

$$R_j = {}^gR_p^T {}^gR_d \quad (1)$$

R_j est une matrice de rotation. En utilisant une convention standard, telle que la convention angulaire de Cardan, la matrice de rotation peut être représentée par trois angles de rotation $R_j = \{[(R_{j\gamma})R_{j\alpha}]R_{j\beta}\}$ - Premièrement, autour de l'axe d_Z , deuxièmement, autour de l'axe d_X , et enfin autour de l'axe d_Y . Ces trois angles peuvent être interprétés comme les angles d'flexion-extension, d'abduction-adduction et de rotation interne-externe de l'articulation. La cinématique de l'articulation du genou est illustrée dans la figure 1.5.

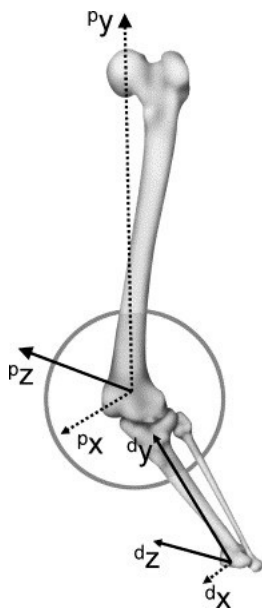


FIGURE 2: Fémur, ${}^gR_p = [p_X, p_Y, p_Z]$ et tibia, ${}^gR_d = [d_X, d_Y, d_Z]$. La cinématique de l'articulation du genou ($R_j = {}^gR_p^T {}^gR_d$) est décrite en utilisant la convention angulaire de Cardan : flexion-extension du genou comme rotation autour de l'axe (d_Z), abduction-adduction du genou comme rotation autour de l'axe (d_X) et rotation interne-externe du genou comme rotation autour de l'axe (d_Y). Reproduit de (Cappozzo et al., 2005)

La figure 1.6 présente un graphique de marche standard pour l'angle d'flexion-extension du genou. Les données sont normalisées dans le temps à un cycle de marche. Les couleurs bleue et rouge représentent les données des côtés droit et gauche, respectivement. L'instant de lever des orteils est indiqué par une ligne verticale sur la hauteur du graphique. L'instant de lever des orteils ou du contact du talon opposé est indiqué par une coche en haut du graphique. La zone grise représente la (moyenne $\pm$ écart type) des données dans une population de référence (Baker et al., 2018).

Paramètres spatiotemporels de la marche

La vitesse de la marche, la longueur de la foulée, la longueur du pas, la largeur du pas, le temps de pas, le temps de phase oscillante, le temps de phase d'appui et la cadence font partie

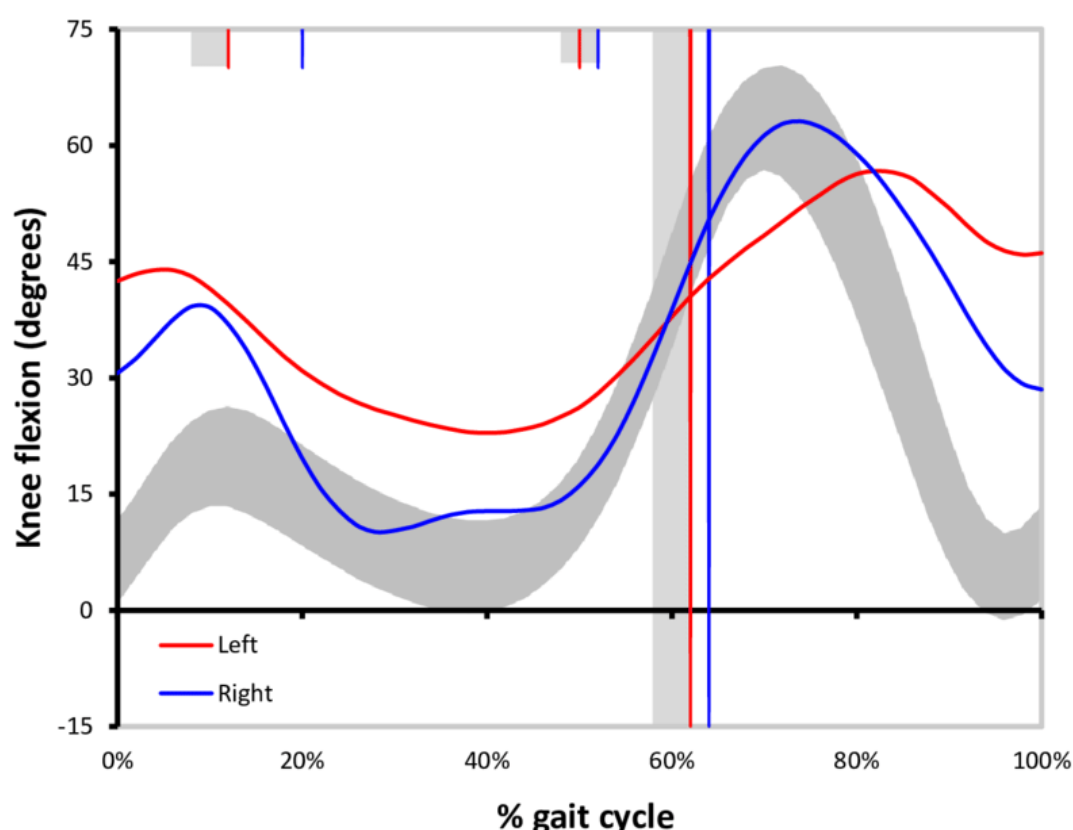


FIGURE 3: Un graphique de marche standard pour la flexion-extension du genou (Baker et al., 2018)

des paramètres spatiotemporels de la marche. Ces paramètres sont calculés sur la base des événements de la marche - par exemple, l'impact du talon et le coup de pied. Les événements de la marche sont illustrés dans la figure 1.1. Le système de référence pour la détection des événements de marche est la plate-forme de force. Le contact du talon est l'instant où la force de réaction verticale du sol est supérieure à un certain seuil. De même, l'instant de lever des orteils est l'instant où la force verticale est inférieure à un seuil.

Intérêt clinique des paramètres de la marche

Dans (Benedetti et al., 2017), les auteurs illustrent qu'il existe plusieurs conditions cliniques pour lesquelles l'AQM comme outil de diagnostic et de pronostic est pertinente. Un article de synthèse (Wren et al., 2011), basé sur 11 articles, a examiné l'effet de l'AQM sur la prise de décision clinique et le traitement. Il a montré que l'ajout de l'AQM aux données d'examen a entraîné des changements dans les plans de traitement pour 52-89% des patients et 41-51% des procédures. En particulier, 37-39% des opérations prévues n'ont pas été réalisées, et 28-40% des opérations réalisées n'ont pas été planifiées avant l'AQM. En outre, l'AQM pourrait également être utilisée pour l'évaluation du risque de chute. Environ un tiers des personnes de plus de 65 ans font des chutes au moins une fois par an (Lord, Sherrington, and Menz, 2001). De nombreuses études, dont (Hausdorff, Rios, and Edelberg, 2001; Hamacher et al., 2011; Paterson, Hill, and Lythgo, 2011; Toebe et al., 2012), ont montré que l'AQM peut être prometteuse pour évaluer le risque de chute et identifier les futurs chutes dans la population des personnes âgées. Par exemple, (Hamacher et al., 2011), à travers une revue systématique, a démontré que la variabilité des **paramètre spatiotemporels** (par exemple, le temps de position

ou le temps de balancement) pourrait être mesuré pour faire la distinction entre les probables chuteurs et les probables non-chuteurs.

Fiabilité des paramètres cinématiques de la marche

Les documents de synthèse (Chiari et al., 2005; Leardini et al., 2005; Della Croce et al., 2005) expliquent que les sources d'erreur peuvent être attribuées à trois éléments principaux : **erreur instrumentale**, **artefact de tissu mou**, et **détermination des repères anatomiques**. L'erreur instrumentale pour le positionnement des marqueurs (généralement < 0,5 mm) a des effets mineurs sur la cinématique des articulations par rapport aux autres sources d'erreur (Benedetti et al., 2017). Le mouvement relatif entre les marqueurs placés sur la peau du corps et l'os sous-jacent est appelé **artefact de tissu mou (STA)**. Il influence considérablement l'estimation de la cinématique des articulations et il est considéré comme la source d'erreur la plus importante dans l'AQM (Leardini et al., 2005). L'identification des points anatomiques et ensuite la **détermination des repères anatomiques** est une autre source d'erreur. Les points anatomiques sous-cutanées sont des surfaces recouvertes de tissus mous ; compte tenu des différentes procédures de palpation utilisées par les médecins pour identifier les points anatomiques, nous comprenons pourquoi la détermination des repères anatomiques est sujette à erreur.

Il existe diverses méthodes et techniques pour limiter la propagation de ces erreurs à la cinématique des articulations. Par exemple, le système EOS® (EOS Imaging, France) peut aider à reconstruire la morphologie des segments osseux à partir d'images radiographiques biplanes à faible dose d'irradiation et les recaler sur les marqueurs externes. Ceci permet de réduire les erreurs attribuées à la détermination des repères anatomiques. Néanmoins, la fiabilité globale des données d'AQM devrait être étudiée.

(McGinley et al., 2009) a examiné vingt-trois études portant sur la fiabilité inter-sessions ou inter-évaluateurs des données cinématiques de la marche. Le **la fiabilité inter-sessions** fait référence aux déviations induites lorsqu'un médecin refait plusieurs fois les séances d'AQM à la même personne. Le **fiabilité inter-évaluateurs** fait référence aux écarts induits lorsque différents médecins effectuent les séances de marche. La figure 1.8 montre la fiabilité entre les sessions ou entre les évaluateurs des paramètres cinématiques de la marche. (Benedetti et al., 2017), suivant (McGinley et al., 2009) et quelques autres études, affirment que les paramètres de marche cinématique les plus fiables sont : **flexion-extension du hanche**, **flexion-extension du genou**, **flexion-extension de la cheville**, **abduction-adduction du bassin**, et **rotation du bassin**. Par ailleurs, la faible fiabilité de certains paramètres cinématiques de la marche n'est peut-être pas la seule limite des systèmes de capture du mouvement basés sur des marqueurs.

Limitations des laboratoires d'AQM

Malgré la valeur clinique de l'AQM, son utilisation n'est pas encore très répandue. Le **coût de l'équipement** d'un système de capture du mouvement basé sur des marqueurs peut atteindre 300 000 \$. De plus, le salaire du personnel (**coûts de main-d'œuvre**) peut être la partie la plus coûteuse d'une séance d'AQM (Simon, 2004). Les systèmes basés sur des marqueurs nécessitent des **environnements de laboratoire dédiés et contrôlés**. Par exemple, l'intensité lumineuse dans le volume de capture doit être entièrement contrôlée. Par conséquent, les dépenses d'un environnement de laboratoire contrôlé dédié sont les autres éléments coûteux qui interviennent dans le coût final d'une session d'AQM. En outre, une séance d'AQM prend beaucoup de temps. En général, la durée d'une séance d'AQM est en moyenne de 60 minutes. Compte tenu de ces limitations, différents systèmes de capture du mouvement ont été proposés comme systèmes alternatifs pour les systèmes basés sur des marqueurs. Cependant, en raison

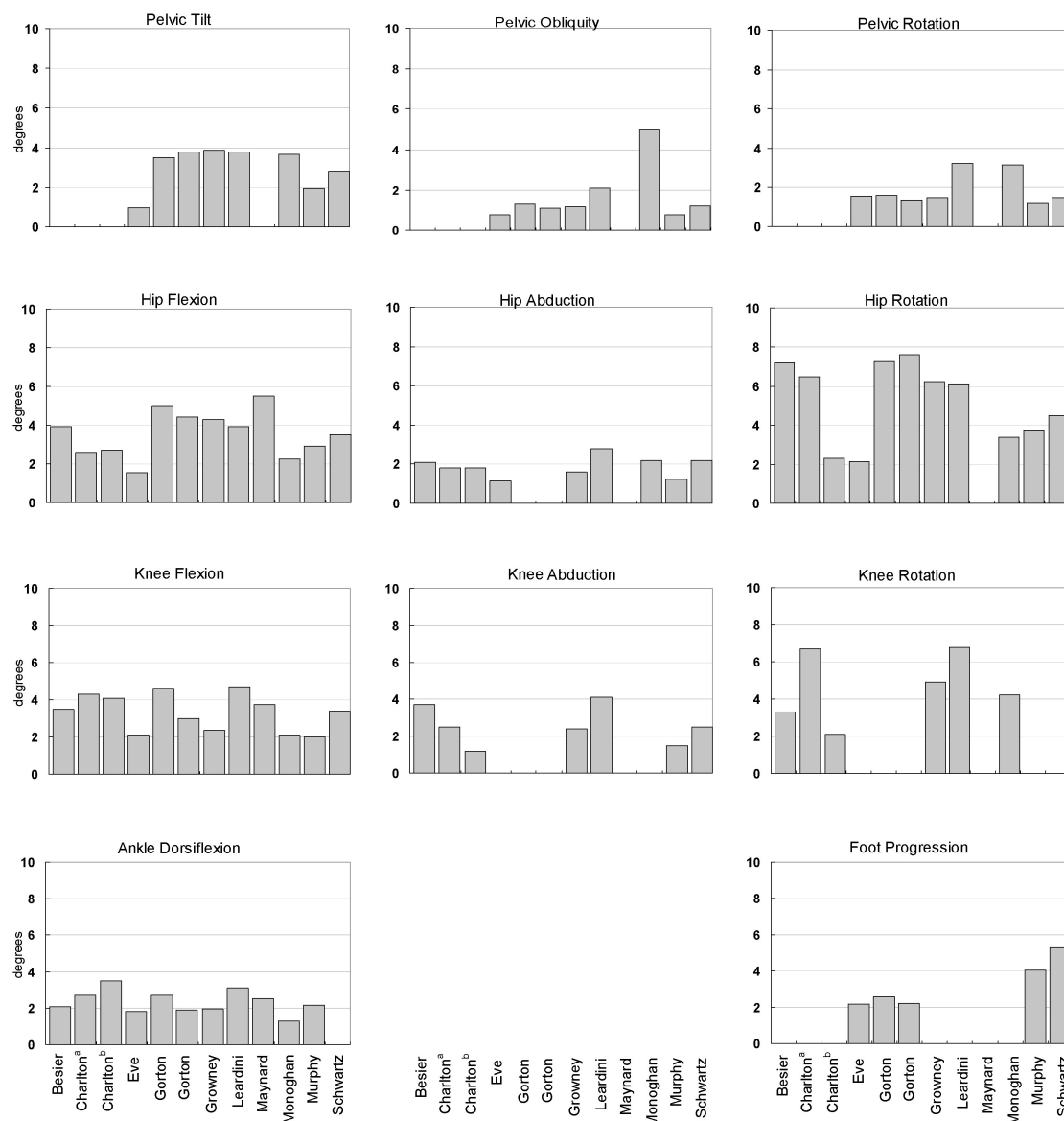


FIGURE 4: Résumé de douze études portant sur la Fiabilité inter-sessions ou inter-évaluateurs des données d'AQM cinématique en tant qu'erreur type ou écart-type (McGinley et al., 2009).

de leur précision et de leur pertinence clinique, les systèmes basés sur des marqueurs sont clairement établis comme la référence. Dans la section suivante, nous étudions les systèmes alternatifs de capture du mouvement proposés.

Systèmes alternatifs de capture du mouvement

Les systèmes de capture du mouvement basés sur les **IMU** (Inertial Measurement Unit, centrale inertielle) et les systèmes **sans marqueurs** font partie des alternatives aux systèmes basés sur les marqueurs. Une centrale inertielle est un dispositif électronique permettant de mesurer des mouvements linéaires et angulaires à l'aide d'un ensemble de gyroscopes et d'accéléromètres. Les systèmes basés sur les IMU, par rapport aux systèmes basés sur des marqueurs, ne nécessitent pas d'environnement de laboratoire et peuvent être utilisés partout. Toutefois, la dérive

inhérente aux estimations de l'orientation (et de la position) est un défaut important de ces systèmes qui influence la précision (Roetenberg, 2006; Schepers, Giuberti, Bellusci, et al., 2018). (Dorschky et al., 2019) présentent une méthode d'estimation des paramètres de la marche du bas du corps à l'aide de 7 capteurs IMU. Les résultats montrent que les erreurs moyennes quadratiques pour les angles du plan sagittal de la hanche, du genou et de la cheville étaient respectivement de 8,2 degrés, 5,5 degrés et 4,3 degrés.

Il existe différents systèmes commerciaux de capture du mouvement sans marqueurs, notamment **Kinetisense**, **wrnchAI**, **DARI Motion**, **The Captury**, **Simi**, et **Theai3D**. Ces systèmes commerciaux utilisent principalement deux types de caméras, soit des caméras de profondeur, soit des caméras RGB. Des méthodes de vision par ordinateur, communément rassemblées sous l'appellation "**Estimation de la Pose Humaine**" (Human Pose Estimation) permettent alors d'obtenir la position des centres articulaires.

Caméras de profondeur (Clark et al., 2019) a examiné la validité des caméras de profondeur pour l'AQM de différentes tâches motrices. Par exemple, pour la marche en surface, il a été signalé que la validité de la caméra Kinect pour la plupart des paramètres cinématiques de la marche était limitée ($r < 0,75$), et que seuls certains paramètres spatiotemporels de la marche tels que la longueur et la largeur des pas étaient valides ($r > 0,75$) (Springer and Yogev Seligmann, 2016). En outre, sur la base de plusieurs études (Pfister et al., 2014; Xu et al., 2015), il existe des erreurs substantielles dans l'évaluation des paramètres cinématiques de la marche sur tapis roulant (moyenne quadratique $> 10^\circ$ ou les limites d'agrément de Bland-Altman $> 10^\circ$).

Caméras RGB Les systèmes sans marqueurs basés sur des caméras RGB calculent les paramètres de la marche selon deux approches différentes. Le processus de travail des systèmes de capture du mouvement sans marqueurs de DARI Motion, The Captury et SIMI Motion consiste à soustraire l'arrière-plan, à estimer les silhouettes du sujet, à ajuster un modèle 3D du corps aux silhouettes pour enfin calculer les paramètres de la marche (voir la figure 1.14). Nous appelons cette approche **Analyse d'image**. D'autre part, les systèmes de capture du mouvement sans marqueurs de wrnchAI et de Theai3D fonctionnent sur la base de l'intelligence artificielle (IA). Nous appelons cette approche **approche basée sur l'IA**.

(Ceseracciu, Sawacha, and Cobelli, 2014) a comparé un système de capture du mouvement sans marqueurs avec un système à base de marqueurs en termes de paramètres de marche cliniquement pertinents. Les différences RMS (moyenne quadratique) entre les systèmes sans marqueurs et les systèmes avec marqueurs étaient de $17,6^\circ$ et de $11,8^\circ$ pour les angles de flexion-extension de la hanche et du genou, respectivement. Dans une étude similaire, (Sandau et al., 2014) a montré que les différences RMS entre les systèmes étaient de $2,6^\circ$ et $3,5^\circ$ pour les angles de flexion-extension de la hanche et du genou, respectivement. Il y avait des différences notables entre les résultats de ces deux études. L'une des raisons sous-jacentes pourrait être les différences techniques des caméras et de tenue vestimentaire des sujets. L'autre raison pourrait être le processus de génération du modèle utilisé dans la seconde étude, selon lequel les centres articulaires et les repères des segments du corps ont été directement transférés du système basé sur des marqueurs au modèle.

Le système commercial de capture du mouvement sans marqueurs de **DARI Motion** est appelé DARI Health. Le nombre de caméras nécessaires est relativement élevé et un environnement de laboratoire dédié est nécessaire. De plus, il n'existe pas de littérature publiée évaluant la précision du système DARI Health ou l'utilisant dans la recherche clinique. (Harsted et al., 2019) a évalué la précision du système de capture du mouvement sans marqueurs **The**



FIGURE 5: Travail avec le système de capture du mouvement sans marqueurs du SIMI Motion. De gauche à droite : enregistrement, extraction de la silhouette, ajustement du modèle 3D du corps humain et calcul des positions et des angles des articulations en 3D. Reproduit à partir de SIMI Motion.

Captury par rapport à un système basé sur des marqueurs. Quatorze enfants (âgés de trois à six ans) ont participé à cette étude. La différence moyenne, les limites inférieure et supérieure d'agrément dans les accroupissements pour l'angle de flexion-extension du genou étaient de 11° , 25° et 4° respectivement. Pour l'angle de flexion-extension de la hanche, ces valeurs étaient respectivement de 1° , 33° et 32° . Bien que les valeurs montrent que la précision du système Captury n'est pas suffisante pour des applications cliniques, il faut noter que la population étudiée n'était composée que d'enfants. De plus, les limites du système Captury en termes de nombre de caméras et d'espace requis sont les mêmes que celles de DARI Health. Le système sans marqueurs du SIMI Motion effectue la capture du mouvement de tout le corps par une série d'étapes, comme le montre la figure 1.14. (Becker and Russ, 2015) a évalué la précision du système de capture du mouvement sans marqueurs du SIMI Motion par rapport à un système à base de marqueurs. Un sujet a effectué vingt-deux types de mouvements. Les écarts types de la différence d'angle pour la flexion-extension de la hanche et du genou étaient de 17° et de 4° . De plus, le nombre de caméras (8 caméras) est relativement élevé.

Approche basée sur l'IA: Les requêtes dans la base de données PubMed avec les mots clés "wrnchAI" et "Theai3D" n'affichent aucun résultat pertinent pour les systèmes de capture du mouvement sans marqueurs. Dans **Theai3D**, la méthode d'estimation de la pose humaine en 3D incorporée est basée sur l'intelligence artificielle, principalement les réseaux neuronaux convolutifs profonds. (Kanko et al., 2020a) a évalué la performance de Theai3D pour la mesure des paramètres spatiotemporels de la marche, et a affirmé que la méthode Bland-Altman ne montrait "aucune différence cliniquement significative" entre les systèmes. Dans une autre étude, (Kanko et al., 2020b) a évalué la fiabilité inter-session et inter-essais des données cinématiques de la marche. Les erreurs inter-session et inter-essais moyennes pour la flexion-extension de la hanche étaient de $2,6^\circ$ et de $2,7^\circ$; pour la flexion-extension du genou, elles étaient de $2,1^\circ$ et de $2,2^\circ$. Cependant, le nombre de caméras est relativement élevé (8 caméras).

L'approche basée sur l'IA a suscité une attention considérable ces dernières années. Par exemple, (Clark et al., 2019), auteur d'une revue de la littérature sur l'utilisation des caméras de profondeur, a affirmé la précision des méthodes d'estimation de la pose humaine, par exemple, **OpenPose** (Cao et al., 2018) et **PoseNet**, pour la mesure des angles d'articulation cinématique est inconnu. La communauté de la vision par ordinateur a beaucoup travaillé sur ce sujet de recherche au cours des dernières années. La figure 1.15 montre le nombre d'études publiées sur l'estimation de la pose humaine 3D basée sur l'IA qui ont évalué leur méthode uniquement sur une base de données d'accès libre. Dans cette figure, la mesure d'évaluation mise en œuvre pour évaluer la précision de l'estimation de la pose humaine 3D est le MPJPE (Erreur de Position Moyenne Par Joint). **L'approche basée sur l'IA fera l'objet de la prochaine section.**

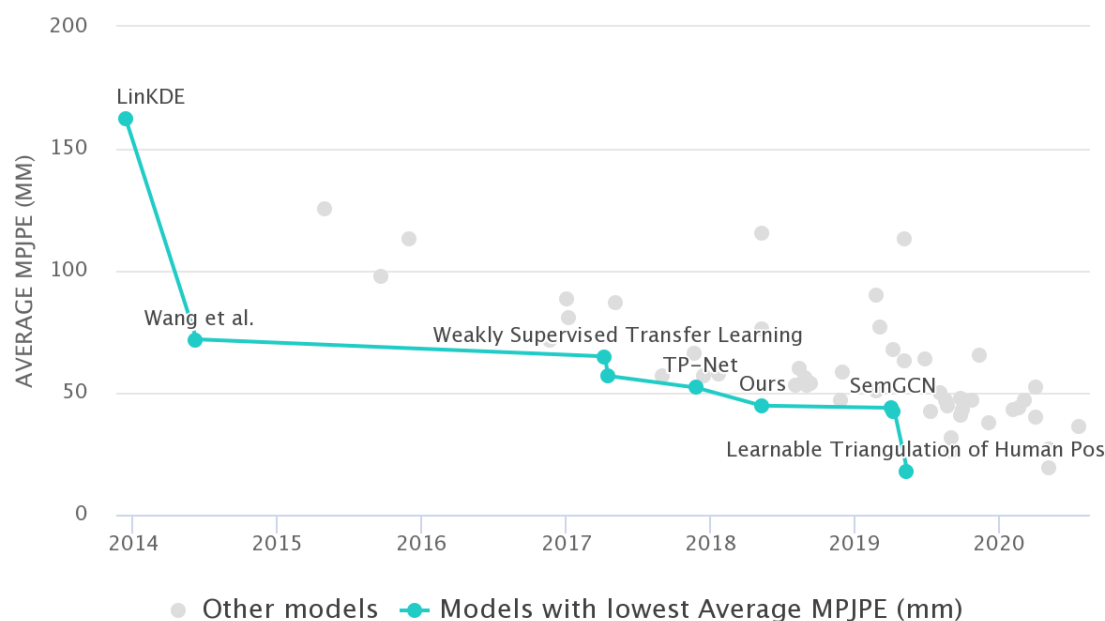


FIGURE 6: Les méthodes d'estimation de la pose humaine 3D basées sur l'IA, évaluées sur un ensemble de données d'accès libre (Human3.6M) (*Human3.6M Benchmark*).

Conclusion

Cette section a démontré le rôle et l'efficacité de l'AQM pour les conditions cliniques et l'évaluation des risques de chute. Ensuite, la procédure de mesure des paramètres de la marche à l'aide du système actuel de référence et la fiabilité des paramètres de la marche ont été étudiées. Les paramètres cinématiques de la marche les plus fiables étaient la flexion-extension de la hanche, du genou, de la cheville, l'abduction-adduction et les angles de rotation du bassin. Les limites des laboratoires cliniques de la marche ont été illustrées, surtout en ce qui concerne les systèmes basés sur des marqueurs. Ainsi, les alternatives aux systèmes basés sur des marqueurs ont été étudiées.

Les systèmes de capture du mouvement basés sur les IMU n'ont pas besoin d'un environnement de laboratoire dédié ; mais ils présentent certaines limites en termes de précision. Les systèmes de capture du mouvement sans marqueurs basés sur des caméras de profondeur sont facilement abordables par rapport aux systèmes avec marqueurs ; mais la précision peut ne pas être adéquate pour les applications cliniques. Les systèmes sans marqueurs basés sur des caméras RGB utilisant l'approche de l'analyse d'image nécessitent un nombre élevé de caméras, un bon contraste du sujet par rapport à l'arrière-plan et un environnement de laboratoire dédié. En outre, aucune étude ne soutient l'adéquation de la précision pour les applications cliniques. L'approche basée sur l'IA a suscité une grande attention ces dernières années. Theai3D, fondé en 2018, un système commercial de capture du mouvement sans marqueurs basé sur l'approche AI, a obtenu des résultats prometteurs. Toutefois, ce système nécessite un nombre élevé de caméras.

De nombreuses études de recherche se sont concentrées sur l'estimation de la pose humaine à l'aide de caméras RGB et de techniques d'intelligence artificielle, obtenant des résultats significatifs ces dernières années. Cependant, leurs performances dans la mesure des paramètres de la marche cliniquement pertinents sont inconnues. Nous avons émis l'hypothèse que les méthodes d'estimation de la pose humaine pourraient fournir un outil puissant pour concevoir et

développer un **système de capture du mouvement sans marqueurs** précis, rentable et facile à utiliser, qui peut être considéré comme une alternative aux systèmes de capture du mouvement basés sur des marqueurs. La section suivante passe en revue les méthodes d'estimation de la pose humaine.

R.2 Estimation de la pose humaine

Les méthodes d'estimation de la pose humaine (Human Pose Estimation, HPE) font référence aux techniques de vision par ordinateur qui peuvent estimer les positions de points clés du corps humain, comme les centres articulaires ou les positions des extrémités des segments. L'objectif de cette section est d'étudier la littérature de l' **estimation de la pose humaine** qui peut aider à concevoir un système de capture du mouvement sans marqueurs pour des applications cliniques.

Base de données de poses humaines

Il existe de nombreuses et diverses bases de données largement utilisées pour l'estimation de la pose humaine - par exemple CMU (*CMU Graphics Lab Motion Capture Database*), HumanEva (Sigal, Balan, and Black, 2010), LSP (Johnson and Everingham, 2011), MPII (Andriluka et al., 2014), COCO dataset (Lin et al., 2014), Human3.6M (Ionescu et al., 2014), CMU Panoptic Studio (Joo et al., 2015), MARCOI (Elhayek et al., 2016), SURREAL (Varol et al., 2017), 3DPW (Marcard et al., 2018). Les bases de données ont des caractéristiques différentes qui les rendent appropriés pour diverses applications. Nous nous concentrons ici sur quatre bases de données, MPII, CMU, HumanEva et Human3.6M. La base de données MPII est largement utilisée pour l'évaluation des méthodes d'estimation de la pose humaine 2D, et les autres sont appropriés pour l'évaluation des méthodes HPE 3D et sont les seuls qui ont utilisé des systèmes de capture du mouvement basés sur des marqueurs pour les annotations.

La base de données MPII (Andriluka et al., 2014) présente 40 522 images de personnes collectées en interrogeant YouTube. La base de données CMU (*CMU Graphics Lab Motion Capture Database*) est l'une des bases de données les plus complètes. Dans cette base de données, il y a 2605 séquences capturées à partir de 109 sujets effectuant un ensemble de mouvements divers. Cependant, pour de nombreuses séquences, seule une vidéo basse résolution capturée par une caméra est disponible, et les données capturées par le système de capture du mouvement basé sur des marqueurs ne sont pas synchronisées avec les vidéos. La base de données HumanEva (Sigal, Balan, and Black, 2009) contient quatre sujets (environ 80 000 images ou 56 séquences) effectuant différents mouvements capturés par plusieurs caméras RGB synchronisées et des systèmes de capture du mouvement basés sur des marqueurs. Les sujets de la base de données HumanEva portaient leurs vêtements habituels. Bien qu'il puisse diminuer la précision du système de capture du mouvement, ce compromis a été accepté pour maintenir le plus possible le réalisme et augmenter la variabilité de l'apparence. L'ensemble de données Human3.6M (Ionescu et al., 2014) contient plus de 3,6 millions de poses humaines 3D diverses. Comme pour HumanEva, des marqueurs ont été fixés sur les vêtements habituels du sujet. Cette base de données, par rapport à HumanEva, a amélioré la variabilité des poses en augmentant les classes de mouvements. Les caractéristiques des bases de données MPII, CMU, HumanEva et Human3.6M sont résumées dans le Tableau 2.1.

HumanEva et Human3.6M sont les seules bases de données dans lesquelles les données de référence, obtenues par des systèmes de capture du mouvement basés sur des marqueurs, sont synchronisées avec les séquences d'images. Cependant, ils ne peuvent pas être utilisés pour l'application spécifique de l'AQM, dont la précision est essentielle. Tout d'abord, les sujets portaient leurs vêtements habituels. La fixation des marqueurs sur les vêtements ordinaires réduirait la précision de l'acquisition de la capture du mouvement. Deuxièmement, le nombre de sujets est faible. HumanEva et Human3.6M contiennent le mouvement de quatre et onze sujets respectivement. De plus, pour Human3.6M, les valeurs de référence pour les centres articulaires ne sont fournies que pour sept sujets sur onze. Cela limite considérablement la

TABLE 1: Caractéristiques du MPII, CMU, HumanEva, et Human3.6M

Base de données	MPII	CMU	HumanEva	Human3.6M
Nombre de sujets	40,522	109	4	7
Type de données	image unique	séquence	séquence	séquence
Annotation	manuel	marqueurs	marqueurs	marqueurs
Données de référence	2D	3D	3D	3D
Vêtements	régulier	serré	régulier	régulier
calibration de caméra	Non	Non	Oui	oui
Données synchronisées	N/A	Non	Oui	Oui
Cas pathologiques	Non	Non	Non	Non

variabilité de la base de données. Par exemple, tous les sujets de Human3.6 étaient de jeunes adultes. Troisièmement, puisque l’objectif de ces bases de données était différent, ils ne contiennent pas le mouvement de cas pathologiques. Ces éléments soulignent la nécessité d’une base de données dédiée sur la pose humaine qui soit appropriée à l’application spécifique de l’étude de la marche.

Métriques d’évaluation

Il existe dans la littérature différentes métriques d’évaluation utilisées pour mesurer la performance des méthodes HPE. Nous en passons ici en revue deux qui sont largement utilisées pour évaluer la performance des méthodes d’estimation de la pose humaine en 2D et 3D.

- **Erreur de Position Moyenne Par Joint (Mean Per Joint Position Error – MPJPE)** (Ionescu et al., 2014) qui est la distance euclidienne moyenne entre les articulations prédites et les valeurs de référence correspondantes.
- **Percentage of Correct Keypoints-head PCKh** ou Pourcentage du Point Clé Correct-Tête (Andriluka et al., 2014) – Pour une pose estimée donnée, une articulation est considérée comme correctement localisée, si la distance euclidienne de l’articulation prédite à la valeur de référence correspondante, est inférieure à 50% de la taille de la boîte de délimitation de la tête.

Ces mesures pourraient ne pas avoir de valeur pratique clinique pour l’AQM. Les paramètres cinématiques qui sont fondamentaux pour l’étude de la marche comprennent, par exemple, l’angle de flexion-extension du genou ou l’angle d’abduction-adduction de la hanche. De plus, la valeur moyenne est un indicateur limité de la probabilité réelle d’erreur, et d’autres paramètres tels que les limites d’agrément Bland-Altman doivent être indiqués.

Modèle de corps humain

Les informations sur la structure cinématique du corps humain, les informations sur la forme et l’apparence sont les principales composantes d’un modèle de corps humain. Ces modèles permettent d’intégrer des connaissances a priori – par exemple, les contraintes cinématiques et dynamiques telles que les limites de mouvement des articulations ou les schémas de mouvement pour des activités spécifiques – dans les méthodes d’estimation de la pose humaine. Les différents types de modèles du corps humain sont présentés dans le tableau 2.13.

Estimation de la pose humaine 2D

L’étude des méthodes HPE 2D est utile car elles constituent la base de nombreuses méthodes HPE 3D. Il existe donc une relation directe entre les performances des méthodes HPE 2D et 3D.

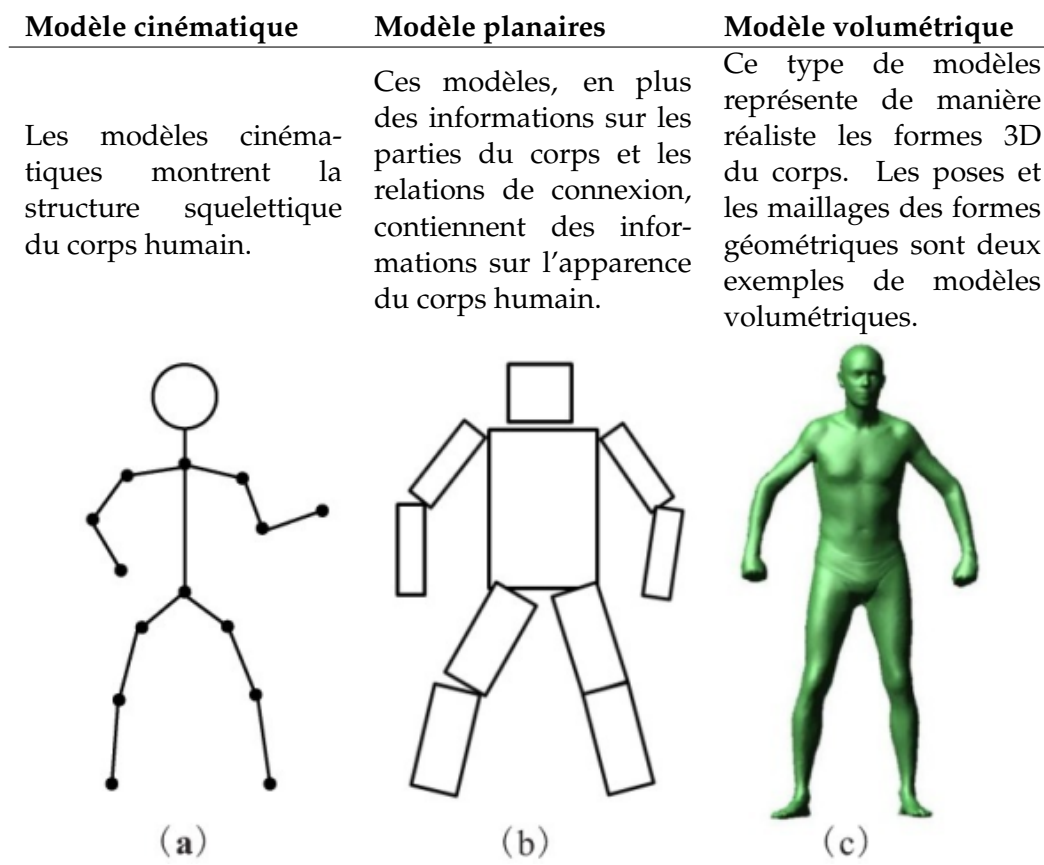


FIGURE 7: Différents types de modèles de corps humain. (a) Modèle cinématique; (b) Modèle planaires; (c) Modèle volumétrique (Gong et al., 2016)

Tableau 2.2 montre les résultats obtenus par les méthodes HPE 2D sur la base de données MPII en utilisant la métrique d'évaluation PCKh.

TABLE 2: Les méthodes HPE 2D, et leurs résultats (%) obtenus sur la base de données MPII en utilisant la métrique d'évaluation PCKh.

Article	tête	épau.	coud.	poign.	hanch.	gen.	chev.	tot.
(Newell, Yang, and Deng, 2016)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
(Zhang, Zhu, and Ye, 2019b)	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
(Ning, Zhang, and He, 2018)	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
(Tang et al., 2018)	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
(Chu et al., 2017)	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
(Liu et al., 2018)	98.4	96.4	92.0	87.8	90.7	88.3	85.3	91.6
(Yang et al., 2017)	98.4	96.5	91.9	88.2	91.1	88.6	85.3	91.8
(Chou, Chien, and Chen, 2018)	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
(Zhu et al., 2019)	98.1	96.7	92.5	88.4	90.8	88.8	95.3	91.8
(Chen et al., 2017)	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
(Ke et al., 2018)	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
(Sun et al., 2019)	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
(Tang, Yu, and Wu, 2018)	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
(Nie et al., 2018)	98.6	96.9	93.0	89.1	91.7	89.0	86.2	92.4
(Zhang et al., 2019a)	98.6	97.0	92.8	88.8	91.7	89.8	86.6	92.5
(Wang et al., 2019d)	98.5	97.1	92.7	88.9	91.6	89.6	86.6	92.5
(Tang and Wu, 2019)	98.7	97.1	93.1	89.4	91.9	90.1	86.7	92.7

Tous les documents, présentés dans le Tableau 2.2, utilisent des réseaux de neurones convolutifs pour apprendre et estimer les différentes poses humaines. L'architecture du réseau pourrait être considérée comme le cœur battant des méthodes. Chaque méthode d'estimation de poses humaines en 2D a son architecture de réseau unique. Cependant, plusieurs architectures de réseau, telles que **VGGNet** (Simonyan and Zisserman, 2014), **Hourglass** (Newell, Yang, and Deng, 2016), et **ResNet** (He et al., 2016), sont utilisées dans diverses méthodes d'estimation de la pose humaine (voir le Tableau 2.3). Dans la section suivante, nous verrons que ces réseaux sont également largement utilisés pour l'estimation de pose humaine en 3D.

TABLE 3: Les méthodes d'estimation de la pose humaine qui ont utilisé **VGGNet**, **ResNet**, et **Hourglass** dans leur structure de réseau.

Architecture de réseau	No.	Références
VGGNet	2	(Nie et al., 2018; Bulat and Tzimiropoulos, 2016)
ResNet	4	(Insafutdinov et al., 2016; Liang et al., 2018; Bulat and Tzimiropoulos, 2016; Zhang et al., 2019a)
Hourglass	14	(Bulat and Tzimiropoulos, 2017; Chu et al., 2017; Yang et al., 2017; Wang et al., 2019d; Zhang et al., 2019a; Tang, Yu, and Wu, 2018; Tang et al., 2018; Zhang, Zhu, and Ye, 2019b; Ke et al., 2018; Tang and Wu, 2019; Zhu et al., 2019; Ning, Zhang, and He, 2018; Chou, Chien, and Chen, 2018; Nie et al., 2018)

Estimation de la pose humaine 3D

Le Tableau 2.4 montre les résultats de l'évaluation des articles sur la base de données Human3.6M. Il montre également l'utilisation d'un modèle de corps humain et le type de données utilisées dans les méthodes d'HPE 3D. Ce tableau comporte deux sections – supérieure et inférieure – qui sont séparées par une ligne horizontale. Dans la partie supérieure, les résultats sont triés en fonction du protocole 1, et dans la partie inférieure, en fonction du protocole 2.

- Protocole 1 : La metrique MPJPE est calculée sans aucun post-traitement supplémentaire sur les poses 3D estimées.
- Protocole 2 : La metrique MPJPE est calculé après le post-traitement – il peut s'agir soit du recalage par rapport à une articulation racine, soit de la méthode Procrustes.

(Iskakov et al., 2019) et (Qiu et al., 2019) ont obtenu des résultats qui font à ce jour référence (MPJPE = 17.7 mm, 26.2 mm) sur la base de données Human3.6M. Il existe des similitudes entre ces deux méthodes. Dans les deux méthodes, le type d'entrée est **une image RGB à vues multiples** ; l'architecture de réseau incorporée est **ResNet-152** ; les **paramètres de la calibration de la caméra** sont utilisés pour récupérer la position 3D des centres articulaires.

TABLE 4: Evaluation de différentes études sur la base de données Human3.6M. **Entrée 1** : une image RGB monoculaire ; **Entrée 2** : une séquence d’images RGB fixes monoculaires ; **Entrée 3** : une image RGB fixe multi-vues ; **Entrée 4** : une séquence d’images RGB multi-vues ; **Modèle** : Modèle de corps humain ; **MPJPE 1** : Erreur (*mm*) – protocole 1 ; **MPJPE 2** : Erreur (*mm*) – protocole 2. La partie supérieure des méthodes du tableau est triée selon le protocole 1 ; la partie inférieure est triée selon le protocole 2.

Article	Entrée 1	Entrée 2	Entrée 3	Entrée 4	Modèle	MPJPE 1	MPJPE 2
(Qiu et al., 2019)	□	□	■	□	■	26.2	
(Gundavarapu et al., 2019)	□	□	■	□	□	32.7	
(Wang et al., 2019c)	■	□	□	□	□	37.6	
(Biswas et al., 2019)	■	□	□	□	■	42.8	
(Chen et al., 2019b)	□	□	■	□	□	46.3	41.6
(Pavlo et al., 2019)	□	■	□	□	□	46.8	36.5
(Kocabas, Karagoz, and Akbas, 2019)	□	□	■	□	□	51.8	45.0
(Rayat Imtiaz Hossain and Little, 2018)	□	■	□	□	□	51.9	42.0
(Liu et al., 2019a)	■	□	□	□	□	52.4	38.4
(Tome et al., 2018)	□	□	■	□	■	52.8	44.6
(Liu, Akhtar, and Mian, 2019)	□	■	□	□	■	54.0	52.3
(Sárándi et al., 2018)	■	□	□	□	□	54.2	
(Núñez et al., 2019)	□	□	□	■	□	54.2	
(Pavlakos, Zhou, and Daniilidis, 2018)	■	□	□	□	□	56.2	41.8
(Wei et al., 2019)	■	□	□	□	□	56.6	42.8
(Sharma et al., 2019)	■	□	□	□	□	58.0	40.9
(Huang et al., 2017)	□	□	□	■	■	58.2	
(Guler and Kokkinos, 2019)	■	□	□	□	■	60.3	46.5
(Liu et al., 2019b)	■	□	□	□	□	61.1	
(Wang et al., 2019b)	□	■	□	□	□	63.7	
(Sun et al., 2018)	■	□	□	□	□	64.1	
(Zhang et al., 2019b)	■	□	□	□	□	66.6	
(Arnab, Doersch, and Zisserman, 2019)	□	■	□	□	■	77.8	54.3
(Jiang et al., 2019)	■	□	□	□	■	87.7	53.8
(Kanazawa et al., 2018)	■	□	□	□	■	88.0	58.1
(Iskakov et al., 2019)	□	□	■	□	□	17.7 ¹	20.8
(Wang et al., 2019a)	■	□	□	□	□		40.7
(Shi et al., 2018)	■	□	□	□	□		43.7
(Kovalenko et al., 2019)	□	■	□	□	■		51.2
(Lee, Lee, and Lee, 2018)	■	□	□	□	□		52.8
(Zhao et al., 2019)	■	□	□	□	□		57.6
(Omran et al., 2018)	■	□	□	□	■		59.9
(Tian et al., 2019)	■	□	□	□	■		62.9
(Zhou et al., 2017)	■	□	□	□	□		64.9
(Chen et al., 2019a)	■	□	□	□	□		68.0
(Pavlakos et al., 2018)	■	□	□	□	■		75.9
(Pavlakos et al., 2019)	■	□	□	□	■		75.9

¹En raison de l’erreur d’annotation dans la base de données Human3.6M, une partie des données d’évaluation a été filtrée.

Conclusion

Ces méthodes d'estimation de la pose humaine basées sur les réseaux neuronaux convolutifs peuvent potentiellement être utilisées comme base pour développer un système de capture du mouvement sans marqueurs, précis, facile à utiliser et rentable pour une application clinique. Toutefois, les bases de données actuelles sur la pose humaine sont insuffisantes pour entraîner et évaluer les méthodes d'estimation de la pose pour l'étude clinique de la marche. En effet, les raisons sont un nombre limité de sujets, l'absence de cas pathologiques et des erreurs introduites par le placement de marqueurs sur les vêtements habituels des sujets. Une nouvelle base de données sur la pose, bien adaptée à l'étude de la marche, devrait être collectée.

La nouvelle base de données devrait être composée de sujets de différentes tranches d'âge, de sujets pathologiques et asymptomatiques. Les valeurs de référence pour les centres articulaires devraient être d'une grande précision. Un système de capture du mouvement basé sur des marqueurs peut être utilisé pour obtenir les valeurs de référence. En effet, les marqueurs doivent être placés directement sur les points anatomiques par des professionnels. De plus, un système d'imagerie médicale (EOS) peut être utilisé pour améliorer la précision des valeurs de référence des centres articulaires. Les caméras RGB (le matériel du système de capture du mouvement sans marqueurs) doivent être calibrées et synchronisées avec le système à base de marqueurs. Dans la section suivante, nous visons à collecter la nouvelle base de données de pose, bien adaptée à l'étude clinique de la marche.

R.3 Base de données ENSAM de poses humaines

Nous souhaitons ici détailler la collecte d'une nouvelle base de données, appelée base de données ENSAM, bien adaptée à l'étude clinique de la marche. Un système de capture du mouvement sans marqueurs sera proposé. Malgré les atouts de la base de données Human3.6M, dans laquelle quatre caméras RGB ont été placées aux quatre coins du volume de capture, la configuration des caméras sera modifiée afin que l'installation des caméras devienne plus simple et plus pratique pour une application en routine clinique.

Matériels et méthodes

Système référence de capture du mouvement basé sur des marqueurs

Le système de capture du mouvement basé sur des marqueurs est le système de capture du mouvement Vicon composé de douze caméras Vicon Vero. La fréquence d'acquisition des données est de 100 Hz. Le système Vicon est équipé d'un dispositif, appelé Lock sync box, qui est utilisé pour synchroniser le système sans marqueurs avec le système Vicon.

Système de capture du mouvement sans marqueurs

Nous avons conçu le système sans marqueurs avec quatre caméras RGB (GoPro Hero 7 Black). Les spécifications choisies pour les caméras sont une résolution de 1920x1080 pixels et une fréquence d'images de 100 images par seconde. La fréquence d'images sélectionnée était la même pour le système basé sur des marqueurs. Nous avons monté chaque couple de caméras sur une seule barre d'aluminium (voir la figure 3.3) et l'avons placé sur la vue frontale et latérale du sujet. Cette configuration de caméra a été choisie pour rendre l'installation et l'étalonnage faciles et rapides. Pour plus de simplicité, comme le montre la figure 3.4, les caméras installées sur la **vision frontale du sujet** seront appelées **caméras frontales**, et celles installées sur la **vision latérale du sujet** seront appelées **caméras latérales**. En moyenne, la distance et l'angle entre deux caméras montées sur une même barre sont respectivement de 75 cm et 15°. Chaque barre en aluminium est montée sur un trépied d'une hauteur de 1 m.



FIGURE 8: **gauche** : la caméra GoPro Hero 7 Black est montée sur la barre d'aluminium à l'aide d'une fixation imprimée en 3D. La position et l'orientation de la caméra peuvent être ajustées par les trois degrés de liberté. **droite** : deux caméras sont montées sur la barre d'aluminium, qui est installée sur un trépied.

Dispositif expérimental

La figure 3.4 montre le plan de l'installation expérimentale. Douze caméras Vicon Vero sont montées sur les murs autour du volume de capture. Les caméras frontales sont placées sur la

vue frontale du sujet, les caméras latérales sur la vue latérale du sujet. Le système de coordonnées global se trouve approximativement au milieu de l'espace de capture. Le sujet se déplace soit vers les caméras frontales, soit dans la direction opposée. Des LED de synchronisation sont placées au sol pour que les caméras frontales et latérales puissent les voir. Le plus long chemin que les système de capture du mouvement sans marqueurs et basé sur des marqueurs peuvent parcourir pour enregistrer le sujet pendant son déplacement est d'environ 4 mètres.

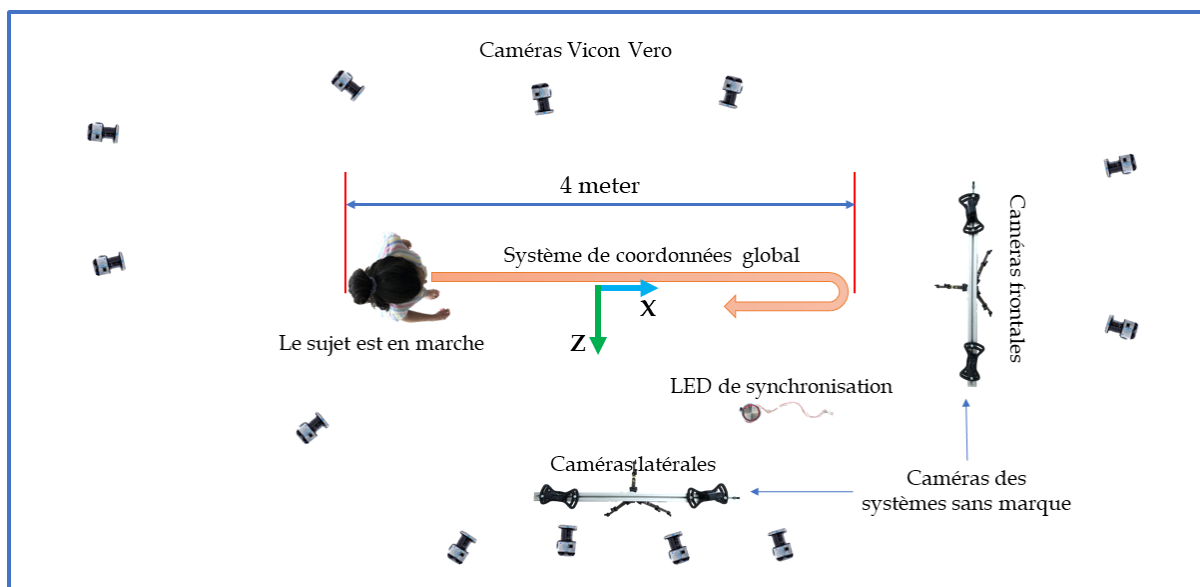


FIGURE 9: Plan du site montrant l'emplacement des caméras du système basé sur des marqueurs (Vicon) et des caméras du système sans marqueurs (GoPro).

Calibration

Le système Vicon est calibré selon les indications du fabricant. En ce qui concerne le système sans marqueurs, le dispositif de la calibration, illustré dans la figure 3.5a, se compose d'un damier (9×14 , 40 mm) et de six marqueurs réfléchissants. Le damier provient de **calib.io**, un producteur commercial de cibles de la calibration. Six marqueurs sont fixés au damier à l'aide de fixations imprimées en 3D. Ensuite, les positions des marqueurs par rapport au damier sont mesurées. À cette fin, une paire de radiographie, comme le montre la figure 3.5b, a été obtenue à l'aide du système EOS[®]. Comme le montre la figure 3.5c, la radiographie a permis de reconstruire en 3D le damier et les marqueurs. Enfin, les positions 3D des marqueurs par rapport au damier ont été calculées.

Une procédure développée a été appliquée pour calibrer simultanément le système de capture du mouvement sans marqueurs et établir le même système de coordonnées globales pour les systèmes avec et sans marqueurs. Pour calibrer le système sans marqueurs, nous appliquons d'abord la procédure développée pour calibrer les caméras frontales, puis les caméras latérales. La procédure de la calibration développée peut être résumée en quatre étapes principales. **Enregistrement des données** : Le dispositif de calibration est placé à vingt positions différentes ($I = 20$). Les caméras frontales du système sans marqueurs enregistrent les vidéos du dispositif de calibration, et le système basé sur les marqueurs enregistre les positions des marqueurs. **Prétraitement des données** : La moyenne des images des vidéos enregistrées est calculée de manière à obtenir une image pour chaque position du dispositif de calibration. **calibration intrinsèque et extrinsèque** : Les paramètres intrinsèques et extrinsèques sont estimés à l'aide de

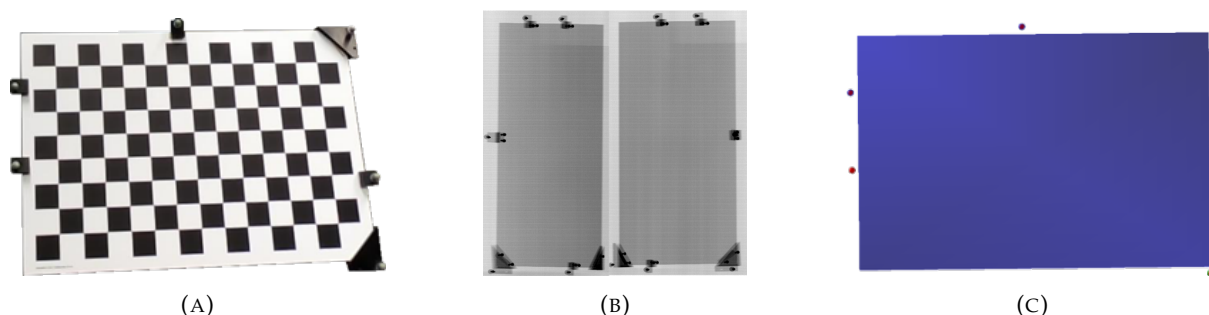


FIGURE 10: (A) Le dispositif de la calibration composé du damier et de six marqueurs. (B) Les radiographies biplanes pour mesurer la position des marqueurs par rapport au damier. (C) Le damier reconstruit en 3D et les marqueurs en utilisant les radiographies biplanes.

l'application MATLAB® (Stereo Camera Calibrator App). **Réglage du même système de coordonnées** : La position 3D des marqueurs par rapport au système de coordonnées des caméras frontales est calculée. Dans un premier temps, les positions 3D des marqueurs ont été enregistrées par le système basé sur des marqueurs. Ensuite, la matrice de rotation R_f et le vecteur de translation T_f entre les caméras frontales du système sans marqueurs et les systèmes de coordonnées du système basé sur des marqueurs peuvent être calculés. La précision de la calibration est évaluée par deux paramètres, reprojection et erreurs 3D. Après la calibration des caméras frontales, la procédure est répétée pour les caméras latérales afin de calibrer l'ensemble du système sans marqueurs.

Synchronisation

Les caméras du système sans marqueurs et le système basé sur des marqueurs enregistrent en utilisant la même fréquence d'acquisition (100 Hz). Par conséquent, la synchronisation vise à trouver le décalage temporel entre les données enregistrées. Une carte Arduino UNO, un microcontrôleur, génère un signal dépendant du temps pour allumer et éteindre un ensemble de LEDs placées dans l'espace de capture. L'état des LEDs – qu'elles soient allumées ou éteintes – dans chaque image des vidéos enregistrées (par le système sans marqueurs) est déterminé par des techniques de traitement d'images (**détection de couleur rouge**). Cette procédure permet de récupérer le signal dépendant du temps. D'autre part, la carte Arduino est connectée au système Vicon en utilisant de Lock sync box et le signal dépendant du temps est directement enregistré par le système Vicon. La corrélation croisée entre les signaux récupérés permet de trouver le décalage temporel entre les données enregistrées.

Collecte de données

Les sujets ont participé à cette étude après avoir donné leur consentement éclairé. Le comité d'éthique compétent a approuvé les travaux (CPP 06036, CPP 06001, Paris VI). Il a été demandé aux sujets d'être en sous-vêtements. Cinquante et un marqueurs réfléchissants ont été placés directement sur les points anatomiques des sujets par des chirurgiens orthopédistes. Pour plus d'informations concernant le placement des marqueurs, voir l'annexe ???. Une paire de radiographies biplanes est ensuite prise avec le système EOS® (EOS Imaging - France) dans une position debout standard (Chaibi et al., 2012). Ensuite, les sujets ont été invités à marcher pendant quinze essais de marche à une vitesse normale de leur choix, tout en étant enregistrés par les systèmes avec et sans marqueurs.

Données d'annotation

Les données d'annotation sont constituées de **la position 3D de 16 centres articulaires ou de segments** (ci-après, pour plus de simplicité, appelés centres articulaires) – tête, cou, épaules, coudes, poignets, tronc, bassin, hanches, genoux et chevilles – et **les boîtes de délimitation des sujets dans les images multi-vues** (les images multi-vues : quatre images enregistrées par les quatre caméras en même temps, instantanées). Les données d'annotation sont obtenues à partir des données enregistrées par le système de capture du mouvement basé sur des marqueurs et le système EOS®.

Une fois que la position des centres articulaires du corps est entièrement déterminée, les centres articulaires sont projetés dans les plans de l'image. Ensuite, les boîtes de délimitation des sujets sont déterminées à l'aide des centres articulaires projetés. La boîte de délimitation est définie par quatre côtés : en haut, en bas, à gauche et à droite. Le côté supérieur de la boîte de délimitation se trouve au-dessus du centre de joint le plus élevé, qui est généralement la tête. Le côté inférieur se trouve sous le centre articulaire le plus bas, généralement une des chevilles. Les côtés gauche et droit sont plus éloignés que les centres articulaires de l'extrême gauche ou de l'extrême droite. Si la boîte de délimitation est de forme rectangulaire, elle doit être transformée en forme carrée, car les entrées des réseaux neuronaux convolutifs sont de forme carrée.

Essais de marche

Plusieurs conditions doivent être remplies pour disposer d'un échantillon valable de la base des données de pose. Premièrement, les sujets doivent se trouver dans l'espace de capture commun des systèmes sans marqueurs et basé sur des marqueurs. Deuxièmement, il ne doit pas y avoir de marqueurs manquants parmi les marqueurs reconstruits d'un essai de marche. Troisièmement, les boîtes de délimitation doivent s'insérer dans les images capturées par les caméras frontales et latérales. Par conséquent, le nombre d'images valides pour former l'ensemble de données de pose dans chaque essai de marche est inférieur aux images capturées par les systèmes sans marqueurs ou basé sur des marqueurs. Dix essais de marche ont été sélectionnés pour constituer la base de données de pose. La sélection a été basée sur la qualité des essais de marche en termes de reconstruction, d'étiquetage et de remplissage des positions manquantes pour les marqueurs dans les essais de marche enregistrés.

Résultats

Calibration

Le système basé sur des marqueurs a été calibré selon les directives du fabricant. Ainsi, la précision de sa calibration n'est pas évaluée. L'erreur 3D moyenne, pour les caméras frontales et latérales du système sans marqueurs, sur l'ensemble des sessions de collecte de données est de 4,3 mm (RMS : 1,0 mm, gamme : 3,0 – 5,8 mm). Il convient de noter qu'une fois que le système sans marqueurs est calibré, et que ses caméras n'ont pas été déplacées, le système sans marqueurs peut enregistrer différentes sessions d'AQM.

Population

Trente et un sujets (19 hommes et 12 femmes) ont participé à cette étude. Les sujets avaient en moyenne 25 ans (écart-type : 9, gamme : 6-44), une taille moyenne de 168 cm (écart-type : 18 cm, gamme : 125 – 199 cm), une masse corporelle de 65 kg (écart-type : 17 kg, gamme :

30-90 kg) et un indice de masse corporelle (IMC) de $22,5 \text{ kg/m}^2$ (écart-type : $3,0 \text{ kg/m}^2$, gamme : $16,9 - 29,4 \text{ kg/m}^2$).

Traitement des données

Le nombre total d'images dans l'ensemble des essais de marche des sujets est de 93331. Figure 3.15 montre un échantillon de la base des données de pose de l'ENSAM – une image multi-vues capturée par les caméras frontales et latérales du système sans marqueurs, seize centres d'articulation du corps projetés sur les plans d'image, et quatre boîtes de délimitation de forme carrée, obtenues à partir du système basé sur des marqueurs.

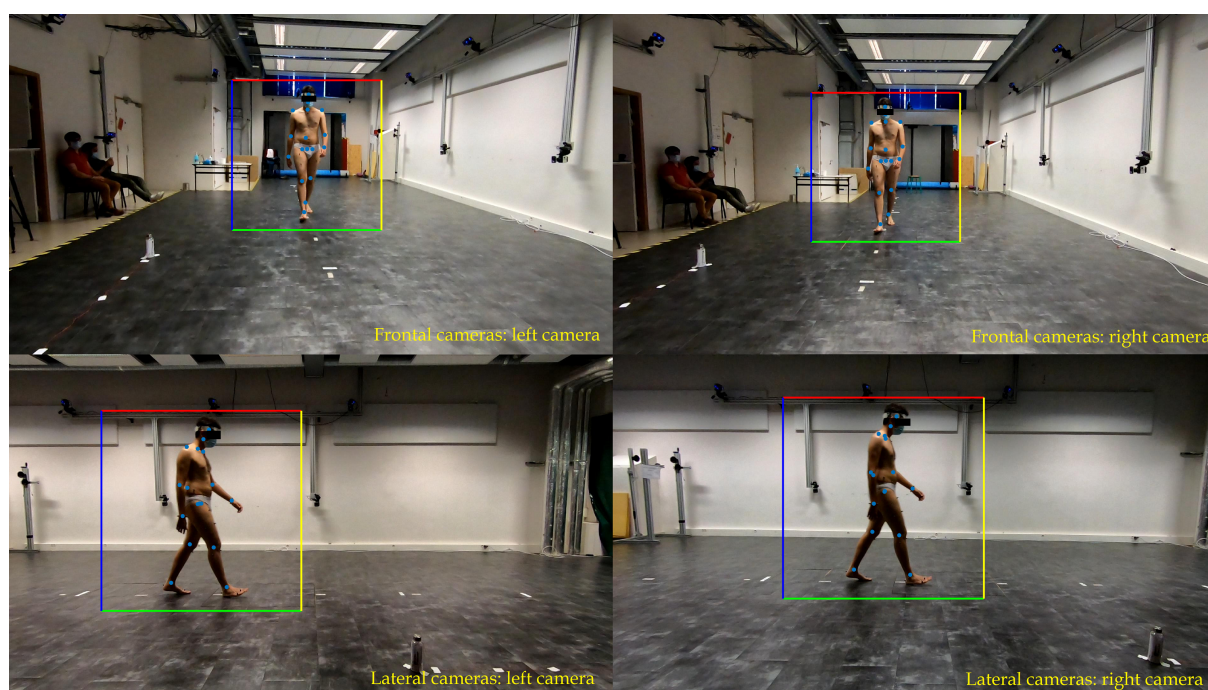


FIGURE 11: La projection est centrée sur les quatre plans d'image des caméras du système sans marqueurs. Les boîtes de délimitation du sujet, de forme carrée, sont récupérées pour chaque plan d'image.

Discussion

La configuration à quatre caméras peut être arrangée de différentes manières. Par exemple, si l'on considère un espace de capture rectangulaire, quatre caméras peuvent être installées aux quatre coins ou sur les côtés. Néanmoins, deux facteurs ont principalement contribué à la conception de l'installation à quatre caméras. Premièrement, l'installation de quatre caméras doit être rapide et simple. Deuxièmement, les caméras doivent être installées sur des vues différentes afin de recueillir plus d'informations. Par conséquent, au lieu de monter chaque caméra sur un seul trépied et de la placer dans une position unique, les deux caméras (paires frontales et latérales) ont été installées sur un trépied à l'aide d'une barre en aluminium.

La procédure de la calibration vise à calibrer le système sans marqueurs et à établir le même système de coordonnées globales que celui du système basé sur des marqueurs. Les paramètres intrinsèques et extrinsèques des caméras frontales ou latérales du système sans marqueurs sont estimés à l'aide de l'application MATLAB Stereo Camera Calibrator App. L'erreur 3D

moyenne, qui représente la précision du réglage du même système de coordonnées globales pour les systèmes basé sur des marqueurs et sans marqueurs, varie de 3,0 à 5,8 *mm*. Même si cette erreur n'est pas restrictive pour la comparaison du système sans marqueurs avec le système basé sur des marqueurs pour la mesure des paramètres de marche cliniquement pertinents, l'erreur pourrait être considérée comme une limitation de cette étude.

Nous avons utilisé un ensemble de LED commandées par un signal dépendant du temps pour synchroniser les caméras du système sans marqueurs. Pour assurer la qualité de la synchronisation, pour plusieurs images de chaque essais de marche enregistrée, comme le montre la figure 3.16, la position des marqueurs capturés par le système basé sur des marqueurs a été projetée dans les plans d'image capturés par le système sans marqueurs. Ensuite, la cohérence entre l'image et les projections de marqueurs a été vérifiée.

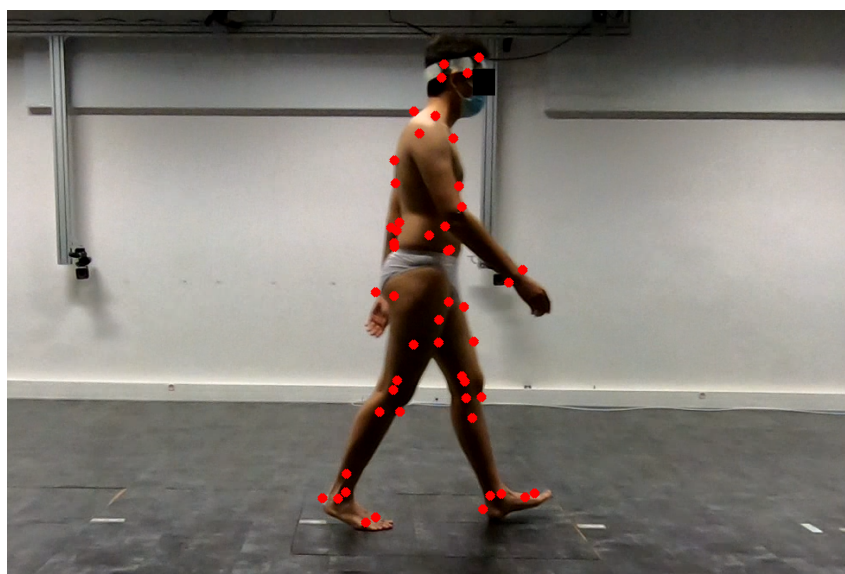


FIGURE 12: Projection de marqueurs (capturés par le système basé sur les marqueurs) sur le plan image d'une caméra latérale du système de capture du mouvement sans marqueurs pour assurer la qualité de la synchronisation.

Conclusion

La base de données de pose de l'ENSAM est conçue et collectée pour évaluer la précision d'un système sans marqueurs ici proposé par rapport aux systèmes basé sur des marqueurs pour l'étude quantitative de la marche en clinique. Le système sans marqueurs est constitué de deux paires de caméras – appelées caméras frontales et caméras latérales – installées sur deux côtés de l'espace de capture. Le système basé sur des marqueurs est un système Vicon équipé de douze caméras Vicon Vero installées sur les quatre côtés de l'espace de capture.

Trente et un volontaires ont participé à la collecte des données de cette étude après avoir donné leur consentement éclairé. Les données d'annotation sont constituées de centres articulaires et de boîtes de délimitation des sujets dans les cadres d'image. La position de seize centres articulaires est extraite à l'aide des systèmes basé sur des marqueurs et d'Eos pour les essais de marche de chaque sujet. Les centres articulaires ont été projetés sur les plans d'image des caméras du système sans marqueurs en utilisant les paramètres de la calibration. Les boîtes de délimitation des sujets de forme carrée ont été obtenues en utilisant les centres articulaires

projetés.

Cette section a décrit tous les éléments de la base des données ENSAM, y compris les objectifs, les dispositifs expérimentaux, la collecte et le traitement des données. Dans la section suivante, nous utilisons la base de données ENSAM pour évaluer la précision du système sans marqueurs par rapport au système basé sur des marqueurs.

R.4 Analyse quantitative de la marche sans marqueurs : validité et fiabilité

Dans cette section, nous allons évaluer le système conçu lors de la collecte base des données de pose de l'ENSAM. La base de données sera divisé en deux ensembles indépendants : l'ensemble d'apprentissage et l'ensemble de test. La méthode d'estimation de la pose humaine du système sans marqueurs est basée sur le réseau de neurones convolutif décrit dans (Iskakov et al., 2019). Ce réseau sera entraîné sur l'ensemble d'apprentissage. L'ensemble de test sera ensuite utilisé pour évaluer la précision du système sans marqueurs par rapport au système avec marqueurs en termes d' **erreur de position des centres articulaires** puis après une détection originale des événements des cycles de la marche, en termes de **paramètres spatiotemporels**, et **paramètres cinématiques**.

Matériels et méthodes

Estimation de la pose humaine

La méthode d'estimation de la pose humaine mise en œuvre (approches de triangulation algébrique et de triangulation volumétrique) a été présentée par (Iskakov et al., 2019). La triangulation algébrique a été utilisée principalement pour estimer le centre du bassin ; ensuite, la triangulation volumétrique a été utilisée pour les autres centres articulaires.

Les méthodes de triangulation algébrique et volumétrique ont été entraînées de bout en bout sur la base de données Human3.6M (Ionescu et al., 2014). Nous avons affiné la méthode sur l'ensemble d'apprentissage de la base de données ENSAM. Cette dernière a été divisée au hasard entre les ensembles d'apprentissage et de test. Les données démographiques des ensembles d'apprentissage et de test sont présentées dans le Tableau 4.1.

TABLE 5: Données démographiques les ensembles de données d'apprentissage et de test ENSAM

	l'ensemble d'apprentissage	l'ensemble de test
Sujets	19 sujets (9 Femmes, 10 Hommes)	12 sujets (3 Femmes, 9 Hommes)
Age (y.)	Mo.: 24 (ET: 8, gamme: 11 – 41)	Mo.: 26 (ET: 10, gamme: 6 – 44)
Taille (cm)	Mo.: 166 (ET: 19, gamme: 125 – 188)	Mo.: 170 (ET: 17, gamme: 133 – 199)
Masse (kg)	Mo.: 62 (ET: 17, gamme: 30 – 86)	Mo.: 68 (ET: 17, gamme: 31 – 90)
IMC (kg/m²)	Mo.: 22.0 (ET: 2.5, gamme: 16.9 – 25.4)	Mo.: 23.3 (ET: 3.5, gamme: 17.5 – 29.3)
Pathologie	4 XLH, 1 Scoliose	3 XLH, 1 Spondylolisthesis

Tout d'abord, nous avons estimé les centres articulaires pour l'ensemble de test, en utilisant la méthode d'estimation de la pose humaine, qui a déjà été entraînée sur Human3.6M. Ensuite, une fois la méthode affinée sur l'ensemble d'apprentissage de la base de données ENSAM, la triangulation algébrique, puis la triangulation volumétrique, ont estimé la position des centres articulaires pour l'ensemble de test. Les erreurs de position des centres articulaires ont été calculées par la distance euclidienne entre les positions articulaires estimées et de référence. Les positions estimées des articulations ont également été utilisées pour estimer les paramètres spatiotemporels et cinématiques de la marche.

Détection des événements de marche et paramètres spatiotemporels

L'algorithme de détection des événements de marche a été adapté de (Zeni Jr, Richards, and Higginson, 2008) et (O'Connor et al., 2007). Cet algorithme a été modifié pour être adapté au système sans marqueurs. Dans cet algorithme, les centres articulaires des chevilles et du bassin ont d'abord été filtrés par un filtre passe-bas (O'Connor et al., 2007). Ensuite, les coordonnées du bassin ont été soustraites des coordonnées des chevilles à chaque image d'essai de marche. Troisièmement, la vitesse des chevilles a été calculée en prenant la dérivée première des coordonnées à l'aide des différences finies. Quatrièmement, les événements de contact du talon et de lever des orteils ont été détectés sur la base de la vitesse relative de la cheville. Au moment du contact du talon, la composante X de la vitesse relative de la cheville est passée de positive à négative ; au moment de lever des orteils, de négative à positive.

Une fois les événements de marche détectés, les paramètres spatiotemporels ont été calculés. Les paramètres spatiotemporels mesurés étaient la longueur de la foulée, la vitesse de la marche, la longueur du pas, la largeur du pas, le temps de pas, le temps de la phase d'appui, le temps de la phase oscillante et la cadence.

Paramètres cinématiques de la marche

Nous distinguons deux types de paramètres cinématiques. Les premiers sont basés uniquement sur les centres articulaires et les seconds sur les repères anatomiques. Les paramètres cinématiques de la marche basés sur les centres articulaires (flexion du tronc, abduction-adduction du fémur, flexion-extension du fémur et flexion-extension du genou basée sur le centre de l'articulation) sont introduits dans le Tableau 4.3.

TABLE 6: Les définitions des paramètres cinématiques de la marche basée sur les centres articulaires

Paremeter	Definition
flexion du tronc	L'angle entre deux vecteurs. Le premier vecteur est formé en reliant le centre du bassin au centre de l'articulation du cou, et le second vecteur est l'axe vertical du système de coordonnées global.
abduction-adduction du fémur	L'angle mesuré dans le plan frontal du laboratoire entre deux vecteurs. Le premier vecteur est formé par la jonction des centres articulaires du genou et de la hanche, projetés dans le plan frontal du laboratoire. Le second vecteur est l'axe vertical du système de coordonnées global.
flexion-extension du fémur	L'angle mesuré dans le plan sagittal du laboratoire entre deux vecteurs. Le premier vecteur est formé par la jonction des centres articulaires du genou et de la hanche, projetés dans le plan sagittal du laboratoire. Le second vecteur est l'axe vertical du système de coordonnées global.
flexion-extension du genou BCA *	L'angle entre deux vecteurs. Le premier vecteur est formé par la jonction des centres articulaires du genou et de la hanche. Le second vecteur est formé par la jonction des centres articulaires du genou et de la cheville.

BCA* : Basée sur le Centre de l'Articulation

Les paramètres cinématiques basés sur les repères anatomiques étaient la flexion-extension de la hanche, l'abduction-adduction de la hanche, la flexion-extension du genou, la rotation du bassin et l'abduction-adduction du bassin. Les paramètres cinématiques de la marche (cinématique des articulations du bassin, de la hanche et du genou) ont été calculés selon la procédure expliquée dans la section 1.1.1. Le système sans marqueurs a estimé deux points pour chaque segment du corps – fémur : centres des articulations de la hanche et du genou, tibia : centres des articulations du genou et de la cheville, bassin : centres des articulations de la hanche. Ainsi, pour chaque segment du corps, un seul axe du système de coordonnées a pu être formé. L'autre axe a été récupéré en utilisant de la connaissance a priori pour former un système de coordonnées.

Fiabilité des paramètres cinématiques de la marche

La fiabilité inter-essai des paramètres cinématiques a été évaluée à l'aide de la méthode proposée par (Schwartz, Trost, and Werver, 2004). Tout d'abord, les paramètres moyens de la marche cinématique ont été calculés sur l'ensemble des essais de marche. Ensuite, les différences des paramètres cinématiques de la marche dans tous les essais de marche par rapport à la moyenne ont été calculées. Ainsi, la fiabilité inter-essai a été obtenue en calculant l'écart-type des différences calculées pour tous les essais de marche.

Résultats

Tout d'abord, l'estimation de la pose humaine a été entraînée sur la base de données Human3.6M et évaluée sur l'ensemble de test de l'ENSAM. Ensuite, la méthode d'estimation de la pose humaine a été affinée sur l'ensemble d'apprentissage de l'ENSAM et évaluée sur l'ensemble de test de l'ENSAM. La différence entre les centres articulaires estimés et les valeurs de référence correspondantes, le long des axes du système de coordonnées global, est présentée dans le Tableau 4.5. Les différences moyennes (biais) étaient élevées (jusque 115,3 mm) et ont considérablement diminué (jusque 7,4 mm) après un affinage sur l'ensemble d'apprentissage de l'ENSAM. De plus, les intervalles de confiance à 95% (2 écarts types) ont été réduits de manière significative pour tous les centres articulaires. Par exemple, le long de l'axe Z pour l'articulation de la hanche, l'intervalle de confiance a été réduit de 106,3 mm à 19,8 mm.

Les erreurs de position (distance euclidienne) sont indiquées dans le tableau 4.6. Il n'y a pas de différence statistiquement significative entre les erreurs de position moyennes des différents sujets ($p=0,95$). De même, il n'y avait pas de différence statistiquement significative entre l'erreur moyenne de position des articulations des sujets pathologiques et des sujets asymptomatiques ($p=0,94$). Sur les seize centres articulaires, les chevilles étaient les plus précises (erreur moyenne = 6,1 mm). Les deuxième et troisième articulations les plus précises étaient les genoux (erreur moyenne = 10,3 mm) et le cou (erreur moyenne = 11,4 mm), respectivement. Le centre articulaire le moins précis était le tronc (erreur moyenne = 26,4 mm).

Les événements cinématiques de la marche ont été déterminés indépendamment en utilisant les centres articulaires estimés par le système sans marqueurs et les centres articulaires de référence du système avec marqueurs. Le décalage maximal entre les événements de la marche obtenu par les deux systèmes était de *deux images* (20 millisecondes) pour tous les événements de la marche. 99 % des différences entre les événements de marche se situaient dans la limite de 1 image (10 millisecondes). Le décalage moyen d'image pour les événements de marche était proche de zéro (moins de 1 milliseconde).

TABLE 7: Erreur de position articulaire (*mm*) dans les essais de marche de tous les sujets le long des axes X, Y et Z du système de coordonnées global.

Axe		chev.	gen.	hanch.	bass.	tronc	cou	tête	poign.	coud.	épau.
Formé sur la base de données Human3.6M											
X	Mo.	3.9	3.9	4.3	6.0	0.5	6.5	0.7	3.8	5.5	5.1
	2ET	68.4	50.4	38.0	23.8	155.0	25.4	42.0	54.8	67.3	42.6
Y	Mo.	25.6	-30.7	-49.4	-50.2	-115.3	-20.4	-14.8	-5.0	-19.4	35.0
	2ET	36.8	47.1	36.6	29.2	132.1	43.9	40.7	37.9	33.0	20.7
Z	Mo.	4.0	3.1	-0.4	0.0	-1.3	1.4	-3.8	5.1	2.7	-0.5
	2ET	30.3	29.3	106.3	11.6	23.9	12.0	14.0	98.7	36.6	75.0
Entraînée sur Human3.6M et affiné sur la base d'apprentissage ENSAM											
X	Mo.	0.0	-0.1	0.1	0.3	-1.1	-0.2	0.1	-0.2	0.5	0.0
	2ET	13.2	16.4	26.3	18.1	29.9	18.4	31.4	27.0	34.5	31.8
Y	Mo.	-0.2	-1.2	-1.9	-3.5	-7.4	1.8	-0.2	1.6	-1.4	0.0
	2ET	6.5	14.1	22.0	22.1	48.1	14.9	2.9	20.9	13.8	15.6
Z	Mo.	-0.2	0.0	0.4	0.3	0.5	0.3	0.1	-1.1	0.5	0.2
	2ET	5.1	8.4	19.8	7.8	13.3	8.7	9.7	24.8	16.9	50.4

TABLE 8: Erreurs de position des centres articulaires (*mm*) des essais de marche de 12 sujets

	chev.	gen.	hanch.	bass.	tronc	cou	tête	poign.	coud.	épau.
Moyenne	6.1	10.3	17.9	14.1	26.4	11.4	14.7	16.0	16.8	21.9
95e percentile	16.0	18.8	31.9	23.2	54.1	23.0	28.4	36.0	35.6	43.2

La différence entre les paramètres spatiotemporels mesurés par les deux systèmes est indiquée dans le Tableau 4.8. Les biais étaient proches de zéro (inférieurs à 0,1). La différence absolue maximale de la vitesse de marche est de 0,014 *m/sec*. La longueur de pas et la longueur de foulée ont une limite supérieure d'agrément similaire (1,506 et 1,437 *cm*) et une limite inférieure d'agrément similaire (1,625 et 1,613 *cm*). Les limites d'agrément pour la largeur de pas sont approximativement (0,713 et 0,720 *cm*) la moitié des limites d'agrément pour la longueur de pas ou la longueur de foulée. La différence absolue maximale de temps de pas, de phase d'appui et oscillante de 20 *millisecondes*. Les limites d'agrément de la cadence sont de -0,04 et 0,04 *pas/seconde*.

Les différences entre les paramètres cinématiques de la marche sont indiquées dans le Tableau 4.9. La différence moyenne pour la flexion du tronc (0,1 degré), la flexion-extension du fémur (0,2 degré), l'abduction-adduction du fémur (0,2 degré) et la flexion-extension du genou basée sur le centre de l'articulation (0,0 degré) était proche de zéro. La différence moyenne pour l'abduction-adduction du bassin (0,5 degré) et l'adduction de la hanche (0,7 degré) était inférieure à un degré. Cependant, la différence moyenne pour la flexion-extension de la hanche (2,6 degrés), la rotation du bassin (3 degrés) et la flexion-extension du genou (3,4 degrés) a montré un biais supérieur à 2,5 degrés. L'amplitude de mouvement a été calculée en moyenne sur l'ensemble des essais de marche des sujets.

La fiabilité des paramètres cinématiques de la marche obtenus par les systèmes sans marqueurs et basé sur des marqueurs est indiquée dans le Tableau 4.10. La fiabilité des paramètres cinématiques de la marche était similaires pour les deux systèmes.

TABLE 9: Différences entre les paramètres de marche spatiotemporels obtenus à partir des systèmes sans et avec marqueurs.

Paramètre	DM ¹	ET ²	LiDA ³	LSdA ⁴	DAM ⁵	RMS ⁶	P ⁷
Vitesse de marche (<i>m/s</i>)	0.001	0.003	-0.007	0.006	0.014	0.003	0.94
Longueur de la foulée (<i>cm</i>)	-0.088	0.778	-1.613	1.437	2.071	0.782	0.93
Longueur de pas (<i>cm</i>)	-0.060	0.799	-1.625	1.506	2.651	0.800	0.90
Largeur de pas (<i>cm</i>)	0.004	0.366	-0.713	0.720	1.355	0.365	0.99
Temps de pas (<i>sec</i>)	0.000	0.006	-0.013	0.012	0.020	0.006	0.94
Temps d'appui (<i>sec</i>)	-0.001	0.008	-0.016	0.014	0.020	0.008	0.79
Temps d'oscillante (<i>sec</i>)	0.001	0.007	-0.014	0.015	0.020	0.007	0.73
Cadence (<i>stHPE/sec</i>)	0.001	0.021	-0.041	0.042	0.091	0.021	0.94

¹Différence Moyenne, ²Écart type, ³Limite Inférieure de l'Agrément Bland-Altman, ⁴Limite Supérieure de l'Agrément Bland-Altman, ⁵Différence Absolue Maximale, ⁶Quadratique Moyenne Différence (Root Mean Square – RMS), ⁷Two-samples t-test p-values.

TABLE 10: Différences (en °) entre les paramètres cinématiques de la marche obtenus à partir les systèmes sans et avec marqueurs.

	Paramètre	AM ¹	RMS ²	DAM ³	DM ⁴	ET ⁵	LiDA ⁶	USdA ⁷
Tronc	Flexion	5.8	1.4	1.1	0.1	1.4	-2.6	2.8
Bassin	Ab-Adduction	3.8	2.8	2.1	-0.5	2.8	-5.9	4.9
	Rotation	7.1	4.1	3.4	3.0	2.8	-2.4	8.4
Hanche	Flexion-extension	31.6	8.4	6.6	2.6	7.9	-13.0	18.2
	Ab-Adduction	9.2	3.5	2.5	0.7	3.4	-5.9	7.4
Fémur	Flexion-extension	23.5	1.9	1.5	-0.2	1.9	-3.9	3.6
	Ab-Adduction	5.8	1.3	1.0	-0.2	1.3	-2.8	2.4
Genou	Flexion-extension	57.1	4.5	3.8	-3.4	3.0	-9.3	2.5
	flexion-extension BCA ⁸	56.2	2.5	1.9	0.0	2.5	-4.8	4.8

¹Amplitude du Mouvement, ²Différences Quadratique Moyenne (Root Mean Square – RMS), ³Différence Absolue Maximale, ⁴Différence Moyenne, ⁵Écart Type, ⁶Limite Inférieure de l'Agrément Bland-Altman, ⁷Limite Supérieure de l'Agrément Bland-Altman, ⁸Basé sur le Centre Articulare.

TABLE 11: Fiabilité inter-essai (degré °) des paramètres cinématiques

	Paramètre	système sans marqueurs	système basé sur des marqueurs
Tronc	flexion	1.0	0.9
Bassin	ab-adduction	0.7	0.7
	rotation	1.6	1.7
Hanche	flexion-extension	1.3	1.4
	ab-adduction	1.2	1.1
Fémur	flexion-extension	2.5	2.5
	ab-adduction	1.0	0.9
Genou	flexion-extension	2.2	1.9
	flexion-extension BCA ¹	2.6	2.5

¹Basé sur le Centre Articulare.

Discussion

Estimation de la pose humaine

Le système de capture du mouvement sans marqueurs se composait de quatre caméras RGB synchronisées et de méthodes d'estimation de la pose humaine basées sur l'apprentissage profond. La comparaison des performances de la méthode d'estimation de la pose humaine sur

l'ensemble de test de l'ENSAM, entraînée sur l'ensemble de données Human3.6M, avant et après le affinage sur l'ensemble d'apprentissage de l'ENSAM, a montré une amélioration drastique. Par exemple, pour le centre de l'articulation de la cheville, les intervalles de confiance à 95% le long des axes du système de coordonnées global étaient inférieurs à 68,5 mm. L'affinement de la méthode d'estimation de la pose humaine sur l'ensemble d'apprentissage de l'ENSAM a permis de réduire ces valeurs à moins ou égal à 13,3 mm. Ces améliorations mettent en évidence la nécessité d'une base de données de pose humaine bien adaptée pour entraîner aux méthodes d'estimation de pose humaine qui sont les bases des systèmes de capture du mouvement sans marqueurs.

La précision d'estimation des différents centres articulaires, comme le montre le Tableau 4.6, n'est pas la même. Par exemple, l'erreur de position des chevilles est de 6,1 mm, alors que l'erreur de position du tronc est de 26,4 mm. Plusieurs facteurs peuvent affecter la précision de l'estimation des différents centres articulaires, notamment l'occlusion, la précision des valeurs de référence des centres articulaires. L'occlusion se produit lorsqu'une articulation est cachée par un autre segment du corps ou une autre articulation. La précision des valeurs de référence des centres articulaires peut être interprétée comme la cohérence des valeurs de référence des centres articulaires entre différents sujets. Les valeurs de référence du tronc ont été obtenues à partir du marqueur placé sur la dixième vertèbre thoracique (T10), ou pour l'épaule, à partir des marqueurs placés sur l'acromion. La précision de ces valeurs de référence a été affectée par l'erreur de placement du marqueur. La méthode d'estimation de la pose humaine a été entraînée et évaluée à l'aide de ces valeurs de référence. Par conséquent, moins les valeurs de référence étaient précises, moins les estimations étaient exactes. En effet, les erreurs de position articulaire du tronc et des épaules étaient les plus élevées (voir le Tableau 4.6).

Détection des événements de la marche

Nous avons modifié la méthode proposée par (Zeni Jr, Richards, and Higginson, 2008) pour qu'elle soit applicable aux systèmes sans marqueurs et basé sur des marqueurs. Néanmoins, les modifications appliquées ont eu un effet négligeable sur la performance de la méthode de détection des événements de la marche. L'algorithme modifié (mis en œuvre dans cette étude) a été comparé à l'algorithme original (Zeni Jr, Richards, and Higginson, 2008) pour le système de capture du mouvement basé sur des marqueurs. La différence maximale entre les événements de marche était de 2 images (20 millisecondes). 97% des différences étaient inférieures ou égales à 1 image (10 millisecondes). Par conséquent, les différences induites par ces modifications étaient négligeables. Ainsi la méthode modifiée pourrait être utilisée pour détecter les événements de marche pour les systèmes de capture du mouvement avec et sans marqueurs.

(Kanko et al., 2020a) a évalué la précision du système de capture du mouvement sans marqueurs Theai3D pour mesurer les paramètres spatiotemporels. Dans cette étude, l'algorithme implémenté pour la détection des événements de la marche (Zeni Jr, Richards, and Higginson, 2008) est similaire à celui mis en œuvre dans notre étude. Les résultats de l'évaluation de (kanko2020) ont montré que seulement 80 % des différences entre les événements de marche étaient inférieures ou égales à deux images, contrairement à 100 % dans notre étude (voir le Tableau 4.8). Cette comparaison peut montrer que notre système sans marqueurs est plus précis que le système de capture du mouvement sans marqueurs Theai3D en termes de détection des événements de marche.

Paramètres spatiotemporels de la marche

Le Changement Minimum Détectable (Minimum Detectable Change **MDC**) est défini comme la quantité minimale de changement nécessaire pour identifier un véritable changement de performance par rapport à un changement dû à la variabilité naturelle du paramètre ou à une erreur de mesure (Mohandas Nair, George Hornby, and Louis Behrman, 2012). Le MDC des paramètres spatio-temporels, basé sur des systèmes de capture du mouvement à base de marqueurs, est rapporté par plusieurs études pour différentes populations. Les limites inférieure et supérieure d'agrément Bland-Altman de tous les paramètres spatiotemporels (voir le Tableau 4.8) étaient inférieures ou égales aux MDCs signalés. Par exemple, les limites inférieure et supérieure d'agrément pour la vitesse de marche étaient de 0,007 m/s et 0,006 m/s, alors que le plus petit MDC était de 0,10 m/s. Par conséquent, **le système de capture du mouvement sans marqueurs était suffisamment précis pour estimer les paramètres spatiotemporels de la marche, y compris la vitesse de la marche, la longueur de la foulée, la longueur du pas, la largeur du pas, le temps du pas, le temps d'appui, le temps d'oscillante et la cadence.**

Paramètres cinématiques de la marche

La différence RMS maximale entre ces quatre paramètres était de 2,5 degrés pour l'angle de flexion-extension du genou basé sur le centre de l'articulation (BCA), qui avait la plus grande amplitude de mouvement. Les limites de l'agrément Bland-Altman étaient également de -4.8° et 4.8°. (McGinley et al., 2009) indique que les erreurs inférieures à 5° sont "susceptibles d'être considérées comme raisonnables dans les erreurs d'AQM". Par conséquent, si nous supposons que les 5° sont les limites acceptables, **nous pourrions considérer que le système sans marqueurs est suffisamment précis pour estimer les paramètres cinématiques de la marche basés sur le centre de l'articulation, y compris la flexion-extension du genou BCA, l'abduction-adduction du fémur, la flexion-extension du fémur et la flexion du tronc.**

Les limites inférieure et supérieure de l'agrément Bland-Altman pour les paramètres de marche cinématique basés sur le système de coordonnées, y compris l'abduction-adduction du bassin, la rotation, l'abduction-adduction de la hanche, la flexion-extension et la flexion-extension du genou, n'étaient pas conformes à la limite de 5°. Cette incohérence s'est généralement produite parce que l'utilisation d'un axe moyen pour la formation du système de coordonnées des segments osseux sur l'ensemble des sujets a conduit à un biais entre les systèmes de capture du mouvement sans marqueurs et basé sur des marqueurs. Même si **il n'y avait pas de concordance adéquate entre les systèmes de capture du mouvement sans marqueurs et basé sur des marqueurs pour les paramètres de marche cinématique basés sur les systèmes de coordonnées**, la fiabilité inter-essai (voir le Tableau 4.10) a montré que **le système de capture du mouvement sans marqueurs était presque aussi fiable que le système de capture du mouvement basé sur des marqueurs.**

Conclusion

Le système de capture du mouvement sans marqueurs conçu est suffisamment précis en termes de paramètres de marche spatiotemporels. En ce qui concerne les paramètres cinématiques de la marche, l'erreur (limites d'agrément Bland-Altman) des paramètres cinématiques de la marche basés sur les centres articulaires est inférieure à 5° et ils peuvent donc être considérés comme suffisamment précis pour des applications cliniques. L'erreur des paramètres basés sur les systèmes de coordonnées n'était pas inférieure à 5°, alors que la fiabilité de tous les paramètres cinématiques de la marche était conforme aux systèmes de capture du mouvement basés sur des marqueurs.

Conclusion générale et perspectives

La capacité de marche est d'une importance inestimable pour la qualité de vie. Les chutes sont une menace pour la santé et l'indépendance des personnes âgées. L'analyse quantitative de la marche peut aider à identifier les futurs chuteurs dans la population adulte plus âgée et à appliquer des mesures préventives pour réduire le risque de chute. L'AQM peut également être utilisée comme un outil de diagnostic et de pronostic pour plusieurs conditions cliniques (par exemple, la paralysie cérébrale). Les instruments de référence actuels pour l'AQM sont les systèmes de capture du mouvement basés sur des marqueurs. Toutefois, les limites de ces systèmes limitent l'utilisation généralisée de l'AQM en routine clinique. L'objectif de cette étude était de développer un système de capture du mouvement précis, facile à utiliser et rentable qui pourrait contribuer à la généralisation de l'AQM dans les applications cliniques.

Parmi les méthodes existantes, nous nous sommes concentrés sur les méthodes d'estimation de la pose humaine basées sur les images RGB pour développer le système de capture du mouvement. Nous avons proposé un système de capture du mouvement sans marqueurs basé sur quatre caméras RGB et des méthodes d'estimation de la pose humaine basées sur l'apprentissage profond. Comme la performance des méthodes d'estimation de la pose humaine basées sur l'apprentissage en profondeur dépend de la base des données d'apprentissage, une base de données dédiée, bien adaptée à l'étude clinique de la marche, a été collectée. La base de données ENSAM contenait les essais de marche de trente et un sujets asymptomatiques et pathologiques d'un large éventail d'âge. Un soin particulier a été apporté à la précision de la collecte des données. Non seulement le placement précis des marqueurs sur des points anatomiques a été effectué par des chirurgiens orthopédistes, mais un système d'imagerie médicale (EOS) a été utilisé pour un recalage précis entre les centres articulaires et les marqueurs externes.

La base des données de pose de l'ENSAM a été utilisée pour entraîner et évaluer le système de capture du mouvement sans marqueurs proposé. La méthode d'estimation de la pose humaine a été entraînée sur la base de données Human3.6M, puis affinée sur un sous-ensemble de la base de données ENSAM. La performance de ce système a été évaluée sur un ensemble de test en termes d'erreurs de position articulaire selon la métrique introduite par la base de données Human3.6M. En outre, conformément à l'objectif de cette étude, la précision de l'estimation des paramètres spatiotemporels et cinématiques de la marche a été évaluée.

Le système sans marqueurs proposé a obtenu des résultats de pointe en termes d'erreur de position de l'articulation pour la marche (erreur moyenne de position de l'articulation réduite de 19,0 mm (Iskakov et al., 2019) à 15,3 mm). En termes de paramètres de la marche, le résultat obtenu a montré que notre système était plus précis que le système Theai3D, système de capture du mouvement sans marqueurs pour détecter les événements de la marche. En ce qui concerne les paramètres spatiotemporels de la marche, les résultats ont montré que les limites d'agrément Bland-Altman étaient inférieures aux changements minimums détectables signalés par différentes études. Par conséquent, le système conçu est suffisamment précis pour mesurer les paramètres spatiotemporels de la marche en application clinique.

En ce qui concerne les paramètres cinématiques de la marche, les limites de l'agrément Bland-Altman pour les paramètres basés sur les centres articulaires étaient inférieures à 5°, ce qui démontre que les erreurs sont raisonnables pour les applications cliniques. Cependant, étant donné que le système de capture du mouvement sans marqueurs ne pouvait qu'estimer les centres articulaires, plusieurs paramètres cinématiques de la marche tels que la flexion-extension de la hanche ou l'abduction n'ont pas pu être mesurés directement avec la précision

requis pour une application clinique. Des études supplémentaires sont nécessaires pour faire progresser le système sans marqueurs afin de pouvoir mesurer ces paramètres de marche cinématique, et des travaux sont en cours à l'Institut à cette fin.

Une autre limitation est que, même si le nombre de caméras utilisées est souvent inférieur aux systèmes commerciaux, quatre caméras peuvent toujours être considérées comme une limitation, car un nombre inférieur de caméras rendrait le système plus facile à utiliser. Par ailleurs, même si cette étude a permis d'obtenir des résultats prometteurs pour l'analyse clinique de la marche, les résultats sont préliminaires. Le nombre de sujets est encore limité et n'est pas représentatif de l'éventail des sujets avec ou sans troubles musculo-squelettiques (âge ≤ 44 ans, en particulier absence de personnes âgées). Toutefois, le protocole de collecte des données est désormais bien défini et il est facile de procéder à une nouvelle collecte de données. L'amélioration de la base de données peut accroître la précision du système et permettre de progresser vers la réduction du nombre de caméras.

Nous avons déjà testé le système sans marqueurs en environnement clinique. Une session d'AQM a été menée à l'hôpital du Kremlin-Bicêtre (Figure 1), et cette séance préliminaire a démontré la faisabilité de cette nouvelle approche, plus conviviale pour les patients et parfois leurs parents. Des travaux sont toujours en cours pour examiner attentivement la fiabilité des paramètres spatiotemporels et cinématiques de la marche. Ce système est opérationnel pour être utilisé en recherche clinique et pourrait ouvrir la voie à l'utilisation de l'AQM comme routine clinique, améliorant ainsi les soins pour la société.



FIGURE 13: Session d'AQM menée à l'hôpital du Kremlin-Bicêtre à l'aide du système de capture du mouvement sans marqueurs. Quatre images sont capturées par les quatre caméras RGB du système sans marqueurs.

General Introduction

Walking plays an essential role in the quality of human life. Walking consists of a repetitive sequence of limb motion aimed to move the body forward and maintain stability (Perry, 1992). Gait literally means a person's manner of walking. Gait analysis is the measurement and ability evaluation of walking (Davis et al., 1991). In gait analysis, the kinematic and spatiotemporal gait parameters are among the measured parameters.

Various studies support the use of gait analysis for fall risk assessment and prediction of future fallers in the elderly adult population. Among older adults aged over 65 years, more than half of the hospitalizations are fall-related (WHO, 2008). Fall incidences might cause a hip fracture, which leads to consequential adverse effects on life quality and increased financial pressures on society (Williamson et al., 2017). Spatiotemporal gait parameters (e.g., stride length) are of high value for fall risk assessment. Stride length measures the traveled distance between the successive ground contacts of one foot. For instance, one standard deviation increase in the stride variability of stride time increases the likelihood of falling about fivefold (Hausdorff, Rios, and Edelberg, 2001).

Also, gait analysis as a diagnostic and prognostic tool for several clinical conditions (e.g., cerebral palsy and acquired brain injuries) is supported by the published studies (Benedetti et al., 2017). (Wren et al., 2011) demonstrated that the addition of gait analysis to the examination data resulted in changes in the treatment plans for 52-89% of patients and 41-51% of procedures. In particular, 37-39% of planned surgeries were not carried out, and 28-40% performed operations were not planned before gait analysis. Concerning acquired brain injuries, (Fuller et al., 2002) investigated the effect of gait analysis on surgical planning. The addition of gait analysis to the examination data resulted in a change in 64% of surgical. Also, the agreement between the surgeons enhanced from 0.34 to 0.76. Nevertheless, despite the clinical value of gait analysis, several limitations restrain its routine clinical use.

The limitations of the gait analysis laboratories are mostly attributed to the currently established instrumentation, which are marker-based motion capture systems. These systems can accurately retrieve the position of a set of markers placed on the skin surface of the subject's bony segments that help compute the gait parameters. However, the cost of equipment, staff salary, the need for a controlled laboratory environment, and the time-consuming test procedure are among the factors that restrain the wide-spread use of gait analysis in clinics. Because of these limitations, several alternative motion capture systems were proposed.

The Inertial Measurement Unit (IMU)-based motion capture system and marker-less motion capture system were among the proposed alternatives. Even though the IMU-based motion capture systems do not require a laboratory environment, some difficulties in terms of the accuracy of these systems limit their use in clinical applications. Marker-less motion capture systems, based on depth cameras or RGB cameras, estimate human poses using various computer vision techniques (referred to as human pose estimation methods). Nevertheless, no published studies support the validity of marker-less motion capture systems for clinical applications.

The objective of this study is to develop a marker-less motion capture system that may help to wide-spread use of gait analysis in clinics. In Chapter 1, we study the experimental protocols for measuring gait parameters, the role and effectiveness of gait analysis, and the limitations of current gold standard systems comprehensively. Then, we review alternative motion capture systems and discuss their limitations. In Chapter 2, we review state-of-the-art human pose estimation methods that may help develop a marker-less motion capture system for clinical applications. In Chapter 3, we propose a marker-less motion capture system. Then, we collect a dataset well-adapted for gait study to train and test the proposed motion capture system. Finally, in Chapter 4, we evaluate the validity and reliability of the proposed marker-less motion capture system in terms of spatiotemporal and kinematic gait parameters.

Chapter 1

Gait Analysis

Walking plays an essential role in the quality of human life. Gait literally means a person's manner of walking. Gait analysis is the systematic study of walking that may help identify gait abnormalities. Numerous studies support the role and effectiveness of gait analysis in clinical conditions as a diagnostic and prognostic tool. Gait analysis may also be used for fall risk assessment of older adults. Falls and subsequent injuries are major public health issues. Among older adults aged over 65 years, more than half of the hospitalizations are fall-related. Falls may have adverse effects on their quality of life (WHO, 2008).

The current established gold standard systems for gait analysis are marker-based motion capture systems. A set of markers are placed on the subject's body segments. The marker-based system retrieves accurately the three-dimensional position of the markers that help calculate the gait parameters. Nonetheless, the limitations of current gait analysis restrain their routine clinical use. In this chapter, we study the experimental protocols for measuring gait parameters, the role and effectiveness of gait analysis, and the limitations of current gold standard systems. Then, we review alternative motion capture systems and discuss their limitations. The objective of this chapter is to find specifications of a motion capture system that may help to wide-spread use of gait analysis in clinics.

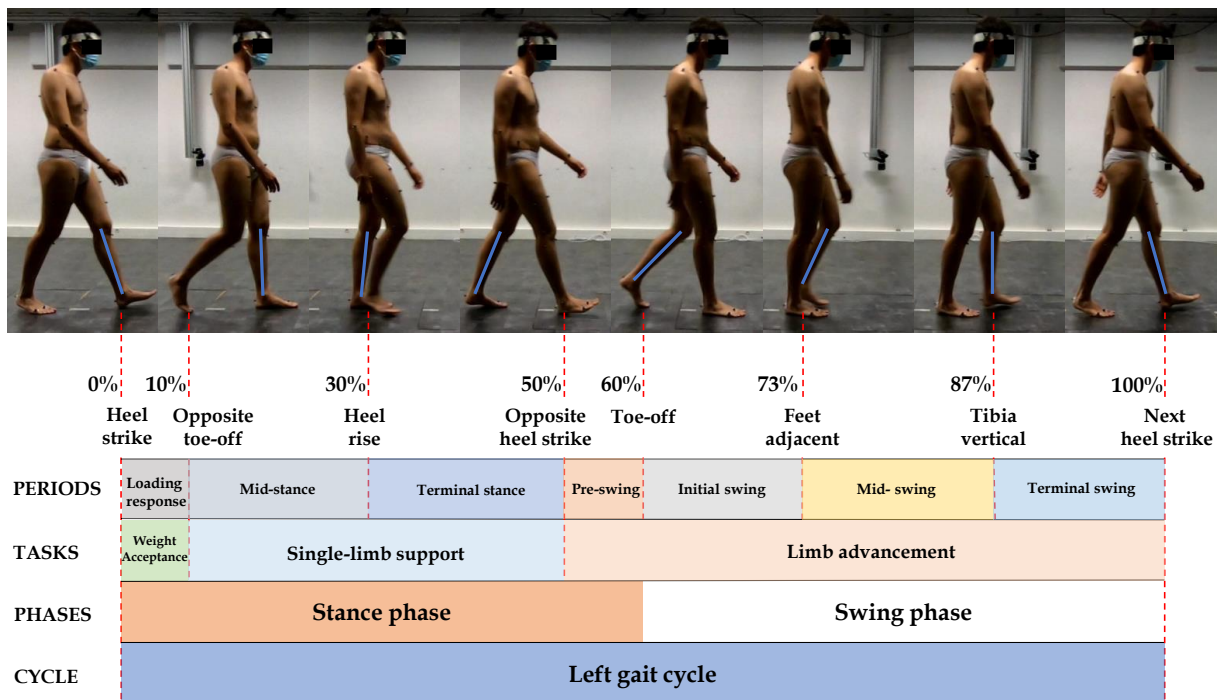


FIGURE 1.1: Divisions of the left gait cycle based on (Kernozek and Willson, 2018)

1.1 Gait analysis

Walking literally means moving at a regular pace by lifting and setting down each foot in turn, in which never both feet are off the ground at once. To this end, walking consists of a repetitious sequence of limb motion with the aim of moving the body forward and maintaining stance stability. In each sequence, two multi-segmented lower limbs, and the total body mass interfere to achieve the walking (Perry, 1992). A gait cycle begins with the heel strike and ends with the successive heel strike of the same foot. The toe-off event divides the gait cycle into stance and swing phases. Stance refers to the entire period during which foot remains in contact with the ground, and it accounts for 60% of a normal gait cycle. On the other side, swing refers to the period during which the foot is in the air or off the ground, and it accounts for 40% of a normal gait cycle. The classical division of a gait cycle into eight phases is illustrated in figure 1.1.

Gait analysis is the measurement and ability evaluation of walking (Davis et al., 1991; Coutts, 1999; Baker, 2006). Gait analysis is developed to address the inadequacy and unreliability of the visual assessment as a clinical skill (Saleh and Murdoch, 1985; Wren et al., 2005). In gait analysis, the **kinematic** and **spatiotemporal gait parameters** are among the measured parameters.

kinematic gait parameters (joint kinematics) represent the relative movement between adjoining bones (Cappozzo et al., 2005). **Spatiotemporal gait parameters** are spatial or temporal descriptors of gait. The current established gold standard systems for clinical gait analysis are **marker-based motion capture systems** and **force platforms**. These systems can help retrieve the gait parameters. In the following section, we explain how the gait parameters are measured.

1.1.1 Kinematic gait parameters

A set of markers are placed on the skin surface of body segments (e.g., tibia, femur, and pelvis) in such a way to estimate their orientation. The three-dimensional position of markers is retrieved by the marker-based motion capture system. For each body segment, two types of coordinate systems are defined, **technical** and **anatomical** frames. The technical frame, as shown in Figure 1.2, formed by three non-aligned markers, aims to record the orientation of the body segment, in three-dimensional space, at every time instant. The location of the technical frame is optional relative to the underlying bony segments and subsequently non-repeatable.

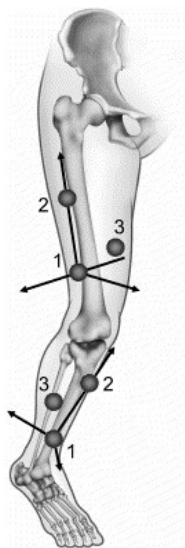


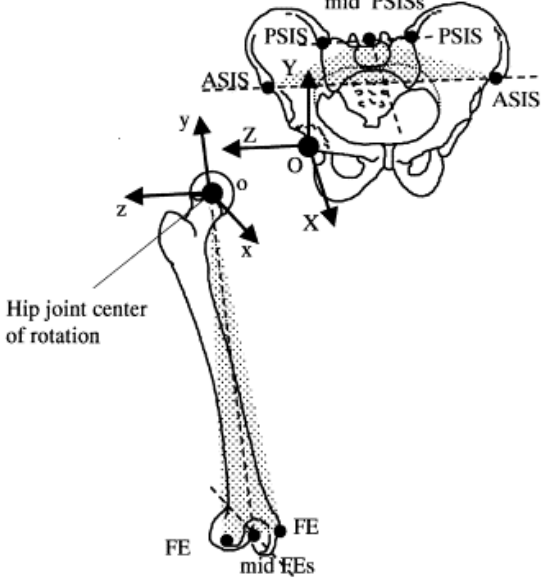
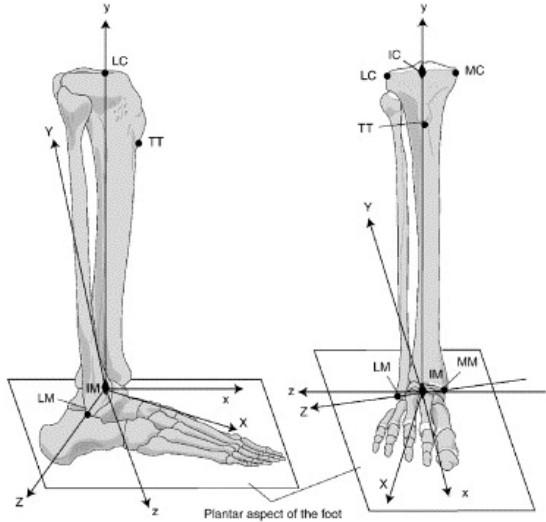
FIGURE 1.2: Marker placement on the right tibia and femur. Three non-aligned markers form the technical frames: p_1 is the three-dimensional position of the frame origin, the Y-axis is defined as $(p_2 - p_1)$, the X-axis as $(p_3 - p_1) \times (p_2 - p_1)$, the Z-axis as the axis which is perpendicular to both the X- and Y- axes and form a right-handed frame.

Adapted from (Cappozzo et al., 2005)

On the other hand, the anatomical frames are defined by the anatomical landmarks, and thereby meet the requirement of inter- and intra-subject repeatability. The definitions of anatomical frames for pelvis, femur, and tibia, recommended by the International Society of Biomechanics (ISB), are introduced in Table 1.1.

There are different techniques to estimate the position of anatomical landmarks and register them to the technical frames. For instance, the anatomical landmarks can be identified by palpation, and their position can be determined by placing markers on them. Also, their position can be determined using a wand equipped with at least two markers. The endpoint

TABLE 1.1: ISB recommendation on definitions of anatomical frames. Images adapted from (Wu et al., 2002)

	
Pelvic and femur anatomical landmarks	Tibia/fibula anatomical landmarks
ASIS: anterior superior iliac spine	MM: medial malleolus
PSIS: posterior superior iliac spine	LM: lateral malleolus
FE: femoral epicondyle	MC: medial tibial condyle
Hip joint center: femoral head	LC: lateral tibial condyle
Pelvic anatomical frame	
Z: line parallel to a line connecting the ASISs, pointing to the right	
X: line parallel to a line lying in the plane containing two ASISs and the midpoint of two PSISs, perpendicular to Z-axis, pointing anteriorly.	
Y: line perpendicular to X- and Z-axis	
O: coincident with the right or left hip joint center	
Femoral anatomical frame	
Y: line connecting the midpoint of FEs and the hip joint center	
Z: line perpendicular to Y-axis, lying in the plane containing the FEs and the hip joint center	
X: line perpendicular to Y- and Z-axis, pointing anteriorly	
O: coincident with the hip joint center	
Tibia/fibula anatomical frame	
Z: line connecting the MM and LM, pointing to the right.	
X: line perpendicular to the plane containing the midpoint of MC and LC landmarks, MM, and LM.	
Y: line perpendicular to X- and Z-axis.	
O: coincident with the midpoint of MM and LM.	

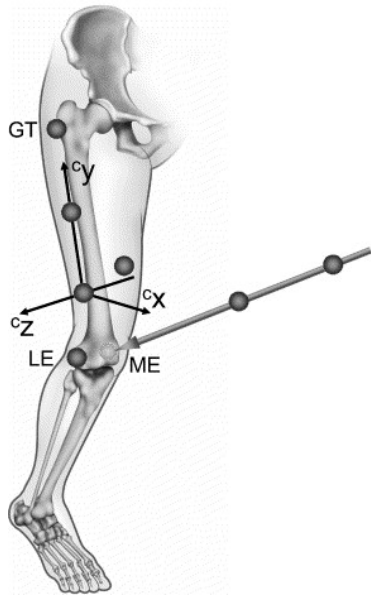


FIGURE 1.3: The femur anatomical landmarks: greater trochanter (GT), medial epicondyle (ME), and lateral epicondyle (LE). The location of the GT and LE, relative to the technical frame $[c_X, c_Y, c_Z]$, are determined by placing two markers on these landmarks. Concerning the ME, the endpoint of a wand equipped with two markers is placed on the landmark. Reproduced from (Cappozzo et al., 2005).

of the wand, which is known accurately relative to the wand markers, is put on the anatomical landmarks and the estimated position of the wand endpoint is registered to the technical frame. These two techniques are illustrated in Figure 1.3.

Meanwhile, some anatomical landmarks (e.g., hip joint center) cannot be identified by palpation. "Predictive," "functional," or medical image-based methods can be utilized. In a predictive method (Bell, Pedersen, and Brand, 1990), the hip joint center is estimated using anatomical assumptions and anthropometric reference data relative to the anatomical landmarks. In a functional method (Leardini et al., 1999), the hip joint center, assumed as the center of rotation of the femur relative to the pelvis, is estimated using marker movement data. There are different medical image-based methods (Kawakami et al., 2005; Sholukha et al., 2006). One of the current methods (Pillet et al., 2014) is based on the low-dose bi-planar X-ray images captured by the EOS[®] system (EOS Imaging – France). The captured images, as shown in Figure 3.9, are

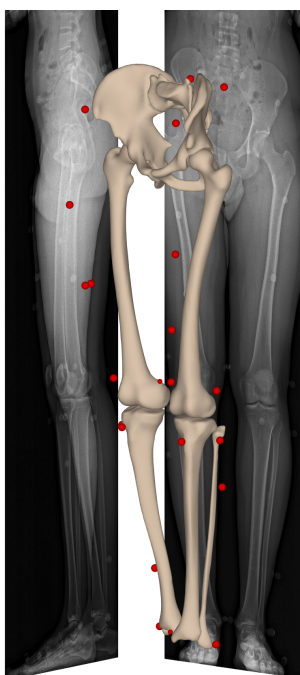


FIGURE 1.4: The bi-planar X-ray images captured by the EOS[®] system; The reconstructed bony segments – pelvis, femurs, tibias, and fibulas; The reconstructed markers are represented by red spheres.

used to reconstruct the 3D models of the bony segments and the markers. Then, the hip joint center can be estimated and registered to the corresponding technical frames. (Pillet et al., 2014) evaluated the accuracy of the EOS system for hip joint center localization. This study showed that the mean accuracy for hip joint center localization was 2.9 mm, while the usual accuracy of non-image-based methods ranges from 13 mm to 30 mm. Therefore, the EOS system helps improve the accuracy of the body joint centers localization.

For joint kinematics description, consider the anatomical frames of two adjacent bony segments, the proximal (p) and distal (d). If ${}^gR_p = [p_X, p_Y, p_Z]$ and ${}^gR_d = [d_X, d_Y, d_Z]$ represent the orientation matrices of the proximal and distal bony segments with respect to the global (g) coordinate system, the joint orientation matrix can be obtained using Equation 1.1 (Cappozzo et al., 2005).

$$R_j = {}^gR_p^T {}^gR_d \quad (1.1)$$

R_j is a rotation matrix that has three degrees of freedom. Using a standard convention, such as Cardan's angular convention, the rotation matrix can be represented by three rotation angles $R_j = \{[(R_{j\gamma})R_{j\alpha}]R_{j\beta}\}$ – First, around the d_Z axis, second, around the d_X axis, and finally around the d_Y axis. These three angles can be interpreted as the joint extension-flexion, abduction-adduction, and internal-external rotation angles. It should be noted that these angles do not describe real rotational movement. Knee joint kinematics is illustrated in Figure 1.5.

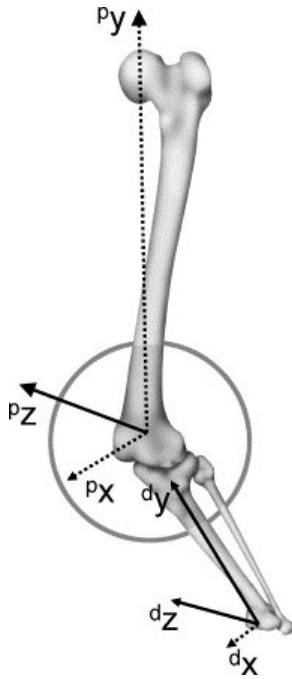


FIGURE 1.5: The proximal (femure, ${}^gR_p = [p_X, p_Y, p_Z]$) and distal (tibia, ${}^gR_d = [d_X, d_Y, d_Z]$) bony segments. Knee joint kinematics ($R_j = {}^gR_p^T {}^gR_d$) described by using the the Cardan's angular convention: knee extension-flexion as the rotation about the (d_Z) axis, knee abduction-adduction as the rotation about the (d_X) axis, and knee internal-external rotation as the rotation about the (d_Y) axis. Reproduced from (Cappozzo et al., 2005)

A standard gait graph for the knee extension-flexion angle is shown in Figure 1.6. In a standard gait graph, the data is time normalized to one gait cycle. Blue (or green) and red colors usually represent the data of the right and left sides, respectively. The toe-off instant is denoted by a vertical line across the height of the graph. The toe-off or heel strike instants of the opposite foot are denoted by a tick mark at either top or bottom of the graph. A gray area represents the (mean $\pm$ standard deviation about the mean) of the data in a reference population (Baker et al., 2018).

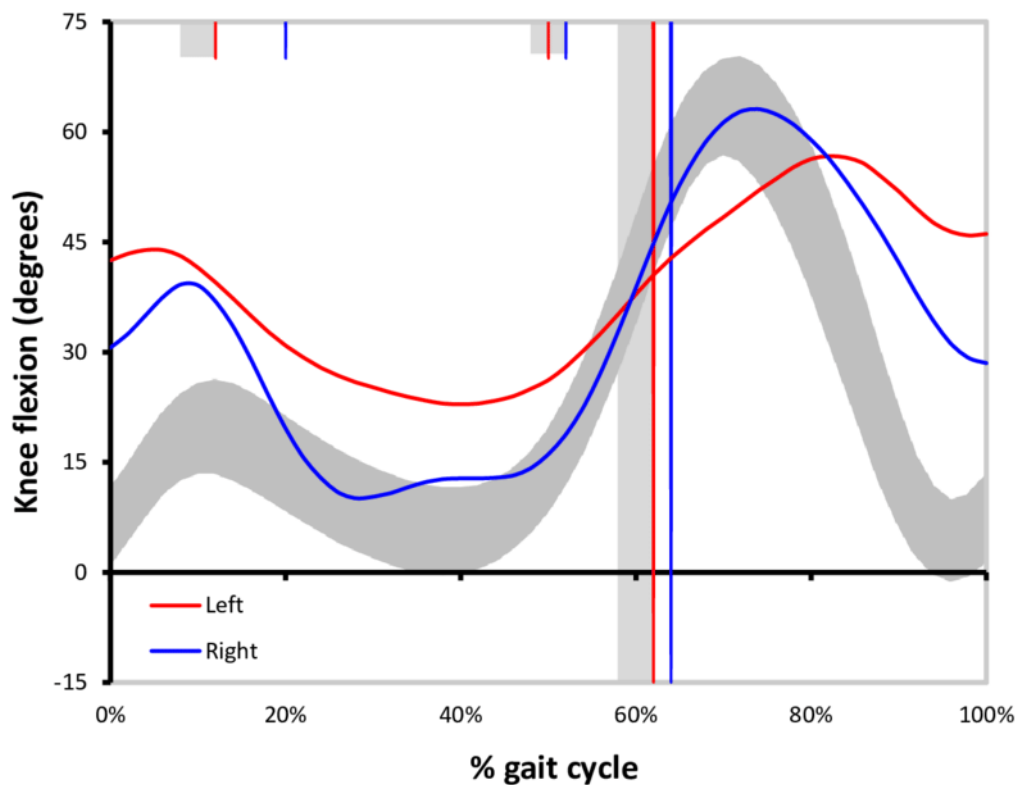


FIGURE 1.6: A standard gait graph for knee extension-flexion (Baker et al., 2018)

A standard gait graph for knee extension-flexion can help examine knee pathologies. Knee excessive knee flexion gait pattern is a gait deviation that can be seen in several clinical conditions (Attias et al., 2019). This gait deviation is identified by higher than normal knee flexion during the stance phase. In cerebral palsy, this pattern is usually attributed to hamstring spasticity or contracture (Presedo, 2018).

1.1.2 Spatiotemporal gait parameters

Gait speed, stride length, step length, step width, step time, stance time, swing time, and cadence are among the spatiotemporal gait parameters. These parameters are calculated based on the gait events – e.g., heel strike and toe-off. The gait events are illustrated in Figure 1.1. The gold standard system for gait event detection is force platforms. The heel strike is the time instant that the vertical ground reaction force is greater than a threshold. Similarly, the toe-off is the time instant that the vertical force is less than the threshold.

Stride length is measured as the distance, along the direction of travel, between the successive points of one foot's heel strikes. Step length is the distance, along the direction of travel, between the point of one foot's heel strike and subsequent heel strike of the opposite foot. Step width is like step length but measured in the perpendicular direction to the axis of travel. The defined parameters are related to displacement, whereas the temporal parameters are related to their time interval. For instance, stance time and swing time represent the time period of the corresponding phases of walking (refer to Figure 1.1). Gait speed is defined as the stride length divided by stride time; Cadence is the number of steps per minute (or second).

Spatiotemporal parameters may be used to assess the risk of falls. Variability of these parameters is associated with the risk of falls in older adults (Yang, 2018). One standard deviation increase in the stride variability of stride time increases the likelihood of falling about fivefold (Hausdorff, Rios, and Edelberg, 2001). Also, the swing time and stride time variability differ significantly between future fallers and non-fallers.

1.1.3 Role of gait analysis

(Baker, 2006) states that there are several reasons to perform gait analysis for clinical conditions; Diagnosis of disease, monitoring of patient's conditions, evaluation and prediction of an intervention (or absence of intervention) outcomes. Numerous studies assessed the role and effectiveness of gait analysis. (Benedetti et al., 2017) illustrates that there are several clinical conditions – **cerebral palsy**¹, **patients with lower limb prostheses**, and **acquired brain injuries** (e.g., stroke) – the published researches of adequate quality support the use of gait analysis as a diagnostic and prognostic tool. Nonetheless, the applicability of gait analysis is not bounded to the mentioned clinical conditions and, for example, could be used for **fall risk assessment**.

Fall risk assessment About one-third of people over the age of 65 experience fall incidences at least once per year (Lord, Sherrington, and Menz, 2001). Hip fracture, because of falls, one of the most consequential, leads to reduced functional ability, adverse effects on life quality, and increased financial pressures² on society (Immonen, 2020).

Numerous studies, including (Hausdorff, Rios, and Edelberg, 2001; Hamacher et al., 2011; Paterson, Hill, and Lythgo, 2011; Toebes et al., 2012), showed that gait analysis might hold promise for assessing fall risk and identifying future fallers in the older adult population. For instance, (Hamacher et al., 2011), through a systematic review, demonstrated that the variability of **spatiotemporal gait parameters** (e.g., stance time or swing time) could be measured to distinguish between fallers and non-fallers.

Cerebral palsy Several studies investigated the effect of gait analysis on decision making and treatment planning. (DeLuca, 1997) reported that the addition of gait analysis to the examination data of ninety-one children diagnosed with cerebral palsy resulted in changes in recommendations for surgery in 52% of the cases. Also, a more recent review paper (Wren et al., 2011), based on 11 articles, reviewed the effect of gait analysis on clinical decision-making and treatment. It showed that the addition of gait analysis to the examination data resulted in changes in the treatment plans for 52-89% of patients and 41-51% of procedures. In particular, 37-39% of planned surgeries were not carried out, and 28-40% performed operations were not planned before gait analysis. On the other hand, certain studies assessed the influence of gait analysis on the outcome. In a study (Wren et al., 2013), two groups of patients were defined — the Gait Report group, where the surgeons received the gait analysis reports and the Control group. The results indicate that gait outcomes showed more improvement when more than half of the gait analysis recommendations were followed.

Acquired brain injuries (Fuller et al., 2002) investigated the effect of gait analysis on surgical planning. The addition of gait analysis to the examination data resulted in a change in 64% of surgical. Also, the agreement between the surgeons enhanced from 0.34 to 0.76. In a more

¹Cerebral palsy is a disorder that is caused by brain damage that happen before or during a baby's birth, or during 3 to 5 years of child's life. Cerebral palsy affects muscle tone, movement, and motor skills.

²(Williamson et al., 2017) estimate the health and social costs, at one year period after the fall incidence, to be about 43,000 \$.

recent study, (Ferrarin et al., 2015) evaluated the effect of gait analysis on clinical decision-making. The treatment plans of 71% of patients were changed after the addition of gait analysis to the patients' examination data.

XLH X-Linked hypophosphatemia (XLH) is an inherited disorder that causes rachitic deformities of the lower limbs and short stature. (Mindler et al., 2020) signified that gait analysis is feasible to quantify disease-specific gait deviations in children with XLH.

1.1.4 Reliability of kinematic gait parameters

To assess the reliability of gait analysis measurements, we should identify the sources of error and learn how they propagate to joint kinematics. The review papers (Chiari et al., 2005; Leardini et al., 2005; Della Croce et al., 2005) explained that the sources of error could be attributed to three main elements: **instrumental error**, **soft tissue artifact**, and **determination of anatomical frames**.

Instrumental errors are of two kinds : systematic and random. The underlying reasons for systematic errors could be either inadequacy of the camera model or inaccurately estimated camera model parameters. The random errors may be attributed to the noise of electronic devices, the transformation of images of markers (captured by infrared cameras) to 2D coordinates, digitization, or marker shape distortion due to velocity effects. Several studies (Chiari et al., 2005; Merriault et al., 2017; Furtado et al., 2019) attempted to assess the accuracy and or precision of the commercial stereophotogrammetric systems. (Merriault et al., 2017) set up two experiments to assess the marker positioning accuracy and precision of the Vicon system (Eight T40S cameras, the capture volume $2 \times 1.5 \times 1 \text{ m}^3$) in static and dynamic conditions. The results showed that the Vicon system positioning performance in static cases is better than the dynamic ones. The mean absolute error in the dynamic test was reported to be less than 0.5 mm. (Benedetti et al., 2017) based on the review papers (Chiari et al., 2005; Leardini et al., 2005; Della Croce et al., 2005), stated that the accuracy and precision for marker positioning (typically $< 0.5 \text{ mm}$) have minor effects on joint kinematics compared to the other sources of error.

The relative movement between the markers placed on the body skin and the underlying bone is called **Soft Tissue Artifact (STA)** that substantially influences the estimation of joint kinematics and is considered the most significant source of error in gait analysis (Leardini et al., 2005). (Reinschmidt et al., 1997) assessed STA's impact by using intra-cortical Hofmann pins with marker triads inserted to the bones of five subjects. The **Root Mean Square (RMS)** difference between the joint kinematics calculated using the skin markers and bone markers for the ankle and knee joints. The RMS difference for knee extension-flexion angle was measured 2.1° . However, for knee abduction-adduction and internal rotation angles, the errors introduced by STA were high compared to the range of motion – more than 60%.

The identification of **Anatomical Landmarks (AL)** — either internal or subcutaneous — and subsequently **determining anatomical frames** is a source of error. The subcutaneous ALs are surfaces covered by soft tissues; considering the different palpation procedure used by physicians to identify ALs, we perceive why the determination of anatomical frames is prone to error. The variability of placing a marker by physicians on an AL (lateral condyle) across six subjects is shown in Figure 1.7. On the other hand, some ALs are internal (e.g., hip joint center) and cannot be identified by palpation. These landmarks can be positioned by a functional (Cappozzo, 1984), an imaging-based (Pillet et al., 2014), or a prediction approach (Bell, Pedersen, and Brand, 1990; Davis et al., 1991). (Della Croce, Cappozzo, and Kerrigan, 1999) investigated

the intra- and inter-examiner precision of ALs position determination. The intra- and inter-examiner precision across the ALs of lower limbs varied from 5 – 21 *mm* and 12 – 25 *mm*.

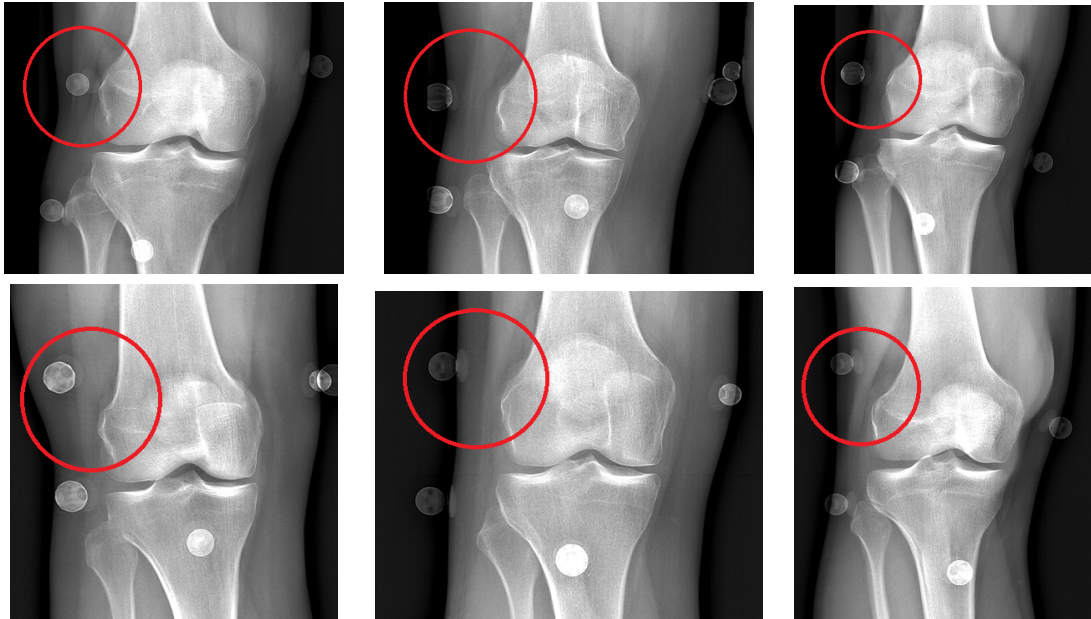


FIGURE 1.7: The red circles show the placement of reflective markers on the lateral condyles of six subjects. The placement of markers is performed by physicians, and the images are captured by the EOS® system.

Indeed, there are various methods and techniques to limit the propagation of these errors to joint kinematics. For instance, the EOS® system (EOS Imaging, France), which could help reconstruct the morphology of bony segments based on low-dose bi-planar X-ray images and register them to the external reflective markers, would reduce errors attributed to the determination of anatomical landmarks. Nonetheless, the overall reliability of gait analysis data should be investigated.

(McGinley et al., 2009), through a systematic review, examined twenty-three studies that addressed inter-session and inter-assessor reliability of gait kinematic data. These twenty-three studies utilized different statistical analyses; Twelve studies reported the reliability of gait kinematic data in terms of standard deviation or standard error of entire kinematic curves or selected kinematic peaks, amplitudes, or events. We review the results of the twelve methods.

The **inter-trial**, **inter-session**, and **inter-assessor reliability** of kinematic gait parameters (Schwartz, Trost, and Wervy, 2004) can be defined as the following. Consider a gait variable (e.g., knee extension-flexion) for a subject, performing a walking trial, in a session, conducted by a physician: The **inter-trial reliability** refers to the stride-to-stride deviations of a subject's gait variable. The **inter-session reliability** refers to the introduced deviations when a physician redoes the gait sessions. The **inter-assessor reliability** refers to the introduced deviations when different physicians perform the gait sessions.

The inter-session or inter-assessor reliability of kinematic gait parameters, as standard deviation or standard error, is shown in Figure 1.8. While (McGinley et al., 2009) state that less than 5° of error can be considered "**reasonable**" in clinical conditions, these results demonstrate that some gait variables, including hip rotation and knee rotation, may exhibit errors above 5°.

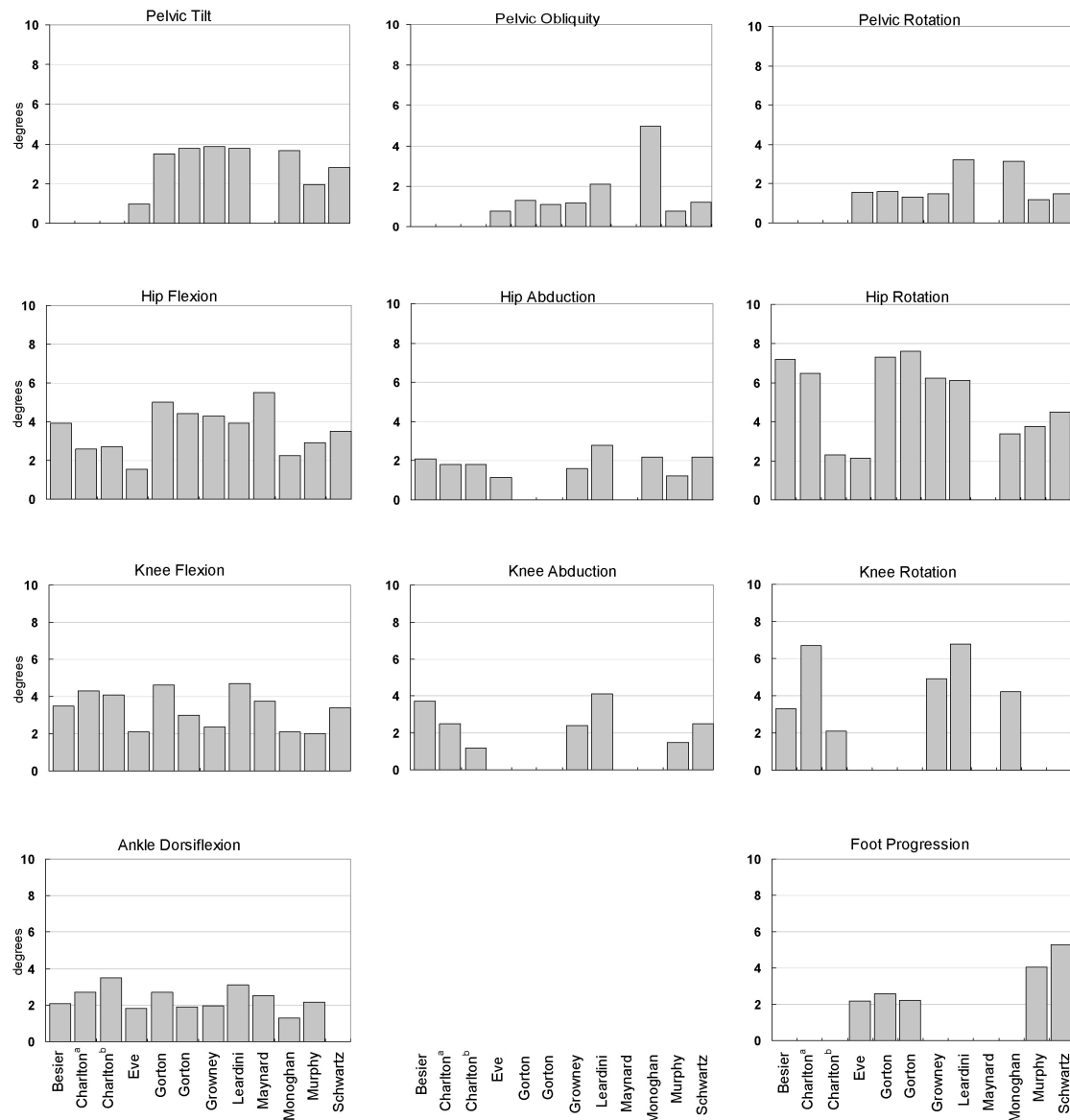


FIGURE 1.8: Summary of twelve studies addressing the inter-session or inter-assessor reliability of kinematic gait analysis data as standard error or standard deviation (McGinley et al., 2009)

Also, even though the error of knee abduction-adduction is less than 5° , it should be considered in comparison to the range of motion (error can be as high as 60% of real motion (Reinschmidt et al., 1997)). (Benedetti et al., 2017), following (McGinley et al., 2009) and a few other studies, affirm that the most reliable kinematic gait parameters are: **hip extension-flexion**, **knee extension-flexion**, **ankle extension-flexion**, **pelvis abduction-adduction**, and **rotation of the pelvis**. Therefore, the interpretation of other kinematic gait parameters (e.g., hip ab-adduction) requires significant consideration because the sources of errors, as mentioned earlier, heavily influence these parameters. Meanwhile, the low reliability of some kinematic gait parameters may not be the only limitation of marker-based motion capture systems.

1.2 Limitations of gait analysis laboratories

The number of published articles supporting the efficacy of gait analysis is increasing substantially over time (Wren et al., 2020; Wren, Chou, and Dreher, 2020). However, despite the clinical value of gait analysis, its usage is not still widespread. The **costs of equipment**, **labor costs**, the need for a **controlled laboratory environment**, and the **time-consuming test procedure** are among the factors that restrain the widespread use of gait analysis for clinical applications.

TABLE 1.2: Comparison of the OptiTrack cameras in terms of price, frame rate, 3D accuracy, Tracking range, FOV, and connection type. The information and images are taken from (Optitrack 2020).



Camera model	Flex13	Prime ^x 13	Prime ^x 13W	Prime ^x 22	Prime ^x 41
Price	599 \$	1,999 \$	2,499 \$	3,499 \$	5,999 \$
Frame rate	100 Hz	240 Hz	240 Hz	360 Hz	180 Hz
3D accuracy*	±0.50 mm	±0.20 mm	±0.30 mm	±0.15 mm	±0.10 mm
Tracking range [•]	6 m	16 m	9 m	21 m	30 m
FOV	46° × 35°	56° × 46°	82° × 70°	79° × 49°	51° × 51°
Connection	USB 2.0	Ethernet	Ethernet	Ethernet	Ethernet

*3D accuracy is for a 9m × 9m area. •Tracking range is estimated using a 14mm marker.

The **Costs of equipment** depend on the configuration — hardware and software — of the marker-based motion capture system. The most costly and primary item is the number and features of cameras — including frame rate, resolution, Field Of View (FOV), and connection type (e.g., USB or Ethernet). The different camera models produced by OptiTrak, their features and prices are presented in Table 1.2. Also, marker-based motion capture software, switch for use with cameras, calibration devices, cables, tripods or fixture for cameras, network card, and a standard PC are the secondary parameters that determine the configuration of the marker-based motion capture system. The configurations of two marker-based motion capture system, designed by using the OptiTrack [online tool](#), are presented in Table 1.3. The first motion capture system is more expensive but requires less setup area than the second system. The other aspects of purchasing a commercial marker-based motion capture system include proprietary software, product warranty, product quality determining its life span, after-sale services, and support that form a commercial product's total price. The price of a marker-based motion capture system can be as high as 300,000 \$.

TABLE 1.3: Configurations of two OptiTrack motion capture systems

	Motion capture system I	Motion capture system II
camera model	Prime ^x 13W	Prime ^x 22
Number of cameras	16 cameras	8 cameras
Capture volume	5m × 5m × 2m	7m × 7m × 2m
Setup area	6m × 6m	9m × 9m
Total price	45,000 \$	35,000 \$

(Benedetti et al., 2017) recommends that the professional profiles for a gait analysis laboratory are: physicians (specialized in relevant areas to the movement science), health care professionals (e.g., physiotherapist), biomedical engineer (specialized in the use of instruments of gait analysis), and human movement scientists. Therefore, the salary of the Staff (**labor costs**) would be the most costly part of a gait analysis session (Simon, 2004).

Marker-based Motion Capture systems require **dedicated controlled laboratory environments**. There are two types of markers for marker-based motion capture systems, active and passive. Active markers emit lights that are tracked by cameras. Passive markers are reflective markers, which are tracked by infrared (IR) cameras. Passive markers do not need a battery and are more common in gait analysis laboratories. The marker-based motion capture systems convert the reflected lights from passive markers to images and then compute markers' position. Nevertheless, The reflections from unwanted sources might be detected as new markers. Also, the ambient light in the capture volume affect the reflections from the passive markers — e.g., the uniform lighting throughout the capture volume help markers appear the same across the whole capture volume (*Optical Motion Capture Guide* 2011). Therefore, the light intensity in the capture volume should be fully controlled. It may clarify why a marker-based motion capture system needs a controlled laboratory environment. Consequently, the expenses of a dedicated controlled laboratory environment is another costly item interfering in the final cost of a gait analysis session.

A gait analysis session is time-consuming. Calibration of the marker-based motion capture system, which should be performed for every gait analysis session, might take 15-30 minutes. The patient preparatory time is about 30 minutes, including the marker placement on the subject's anatomical landmarks. Depending on the conditions, testing may last 5-30 minutes. Generally, the duration of a gait analysis session averages 60 minutes. Moreover, data processing (e.g., marker labeling and filtering) takes nearly 30 minutes. As the duration of a gait study increases, the costs increase.

The marker-based motion capture systems are costly and need dedicated controlled laboratory environments. If the marker-based motion capture system's total price could be reduced or it could be installed easily at different places, it would help improve the widespread usage of gait analysis in clinical applications. Labor costs are high because specific tasks like marker placement on the anatomical landmarks can be performed only by professional individuals. The marker placement is a time-consuming procedure that increases the total costs of a gait study and reduces the cost-efficiency. By considering these limitations, different motion capture systems were proposed as alternative systems for marker-based motion capture systems. However, because of the accuracy and clinical relevance of marker-based motion capture systems, they are clearly established as the gold standard. In the following section, we study the proposed alternative motion capture systems.

1.3 Marker-less and IMU-based Motion capture systems

The **marker-less** and **IMU³-based motion capture systems** are among alternatives for marker-based motion capture systems. In the following, we will investigate these alternative systems and discuss their pros and cons.

³Inertial Measurement Unit

1.3.1 IMU-based motion capture systems

Inertial Measurement Unit (IMU) is an electronic device for measuring linear and angular motions using a set of gyroscopes and accelerometers. IMU-based motion capture systems, in comparison to marker-based motion capture systems, do not require a laboratory environment and can be used anywhere. However, the inherent drift of the orientation (and position) estimates is a significant shortcoming of these systems that influences the accuracy (Roetenberg, 2006; Schepers, Giuberti, Bellusci, et al., 2018).

There are several commercial producers of IMU-based motion capture systems — e.g., **Xsens**, **Shadow**, and **Technaid**. The placement of IMU sensors on the human body is shown in Figure 1.9. The price of the IMU-based motion capture systems depends on the system's configuration. For example, the ballpark price of a typical setup for Xsens MVN Analyze motion capture system is in the range of 40,000 \$. The motion capture system consists of 17 IMU sensors, required setups for receiving and synchronization of the sensors, and proprietary software.



FIGURE 1.9: Placement of IMU sensors on the human body. Reproduced from Xsens.

(Caldas et al., 2017) authored a systematic review of gait analysis using IMU-based motion capture systems. This review compares the accuracy in determining kinematic gait parameters, including joint angles. For instance, (Goulermas et al., 2008) estimate the ankle, knee, and hip joint kinematics using an IMU-based motion capture system and a particular general regression neural network. However, 40% of errors were more than 5°. Also, (Chalmers et al., 2014) compared an IMU-based motion capture system against a marker-based motion capture system for measuring the "foot angle" using an IMU placed on the dorsal part of the shoe. The achieved RMS error was 4.9° in normal walking and 6.5° in a combination of normal and toe walking. In a more recent study, (Dorschky et al., 2019) present a method for estimating the lower body gait parameters using 7 IMU sensors. They formulated an optimal control problem to track the IMU data using a planar human body model. To obtain a unique solution and reduce the noise effects, they minimized the muscular effort. The results show that the RMS errors for hip, knee, and ankle sagittal plane angles were 8.2°, 5.5°, and 4.3°, respectively.

In a very recent review paper, (Weygers et al., 2020) focused on the lower-limb gait analysis using IMU-based motion capture systems. They stated that despite the performed research, there are still difficulties that limit the use of IMU-based motion capture systems in clinical applications. They consider that the inherent drift, sensor-to-segment alignment, and initial sensor orientation are among the complexities and difficulties of using IMU-based motion capture systems. Besides the doubts on these systems' accuracy, even though, on average, IMU-based motion capture systems' costs are less than marker-based systems, based on the stated prices, IMU-based motion capture systems might not be viewed as easily affordable products.

1.3.2 Marker-less motion capture systems

Marker-less motion capture cameras are essentially proposed to be a cost-effective and easy-to-use alternative for the marker-based motion capture systems. There are different commercial marker-less motion capture systems, including **Kinetisense**, **wrnchAI**, **DARI Motion**, **The Capture**, **Simi**, and **Theai3D**. These commercial systems mainly utilize two types of cameras, either depth cameras or RGB cameras. The software is based on the techniques, which are referred to as "**Human Pose Estimation**" by the computer vision community. Three-dimensional (3D) human pose estimation refers to processing 2D images (captured by either depth cameras or RGB cameras) to locate a set of anatomical landmarks or joint centers in the three-dimensional capture volume.

Depth cameras

The **Kinetisense** system (as shown in Figure 1.10) is the only commercial marker-less motion capture system that uses a depth camera — e.g., Microsoft Kinect or Intel D415 — to measure the gait analysis parameters. We queried the PubMed dataset with the keyword "Kinetisense," and no relevant article to the performance evaluation of the system was found. The provided white paper on the Kinetisense's website supports the validity of marker-less motion capture systems. However, the camera setup consists of eight RGB cameras and not depth cameras. Therefore, no published study assessed the validity of the Kinetisense against marker-based motion capture systems, and their use in clinical research is very limited. In the following, we study the literature that attempted to assess the depth cameras as marker-less motion capture systems.

(Schmitz et al., 2014) placed a jig (shown in Figure 1.11) in several static postures. A single Microsoft Kinect camera and a marker-based motion capture system record the object simultaneously. A digital inclinometer measured the ground-truth values for sagittal and frontal joint angles. The marker-based and marker-less motion capture system agreed with the inclinometer by less than 0.5° . This study provided an initial step toward using the Microsoft Kinect camera for measuring clinically relevant gait parameters. However, the angle was only measured for a knee model in several static postures.

In another study, (Bonnechère et al., 2014) assessed the Microsoft Kinect camera's validity and reliability as a marker-less motion capture system against a marker-based motion capture system. Forty-eight healthy adults participated in this study and performed and repeated multiple movements (shown in Figure 1.12) for ten times in two separate sessions. The Microsoft Kinect SDK v1.5 software estimated a skeleton model that was used to estimate the joint angles. The RMS errors for knee extension-flexion were 13° and 14° in the first and second sessions, respectively. The minimum RMS error was for shoulder abduction with 3° and 4° in the two sessions. The results showed that the system's accuracy is not proper for clinical applications.



FIGURE 1.10: The Kinetisense's setup for marker-less motion capture. Reproduced from Kinetisense.

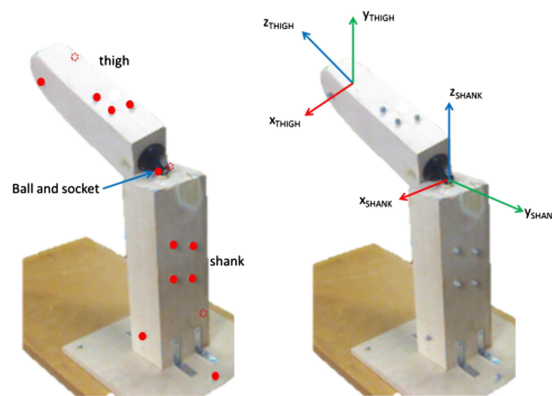


FIGURE 1.11: A jig with a ball-and-socket joint. This knee model was used to compare a Microsoft Kinect camera as a marker-less motion capture system with a marker-based motion capture system. (Schmitz et al., 2014)

(Schmitz et al., 2015) utilized a 3D scanner and a Microsoft Kinect camera to estimate the hip and knee joint angles during a slow squat motion. The joint angles were also measured by a marker-based motion capture system. At each time instant, the 3D scanned segments (pelvis, right femur, and right tibia) were aligned to the Kinect's depth map, and then the joint angles were computed using a set of virtual markers attached to each segment. The 95% CI (Confidence Interval) lower and upper limits of agreements were -2.4° and 6.3° for knee flexion, and -18.9° and 6.0° for hip flexion. The results suggested that the Microsoft Kinect SDK may not be appropriate for measuring joint angles. However, a Kinect camera may hold the potential to be advanced as a marker-less motion capture system. On the other hand, this study only examined the system's validity for a simple movement, and the system's capability during more dynamic tasks like walking is to be assessed.

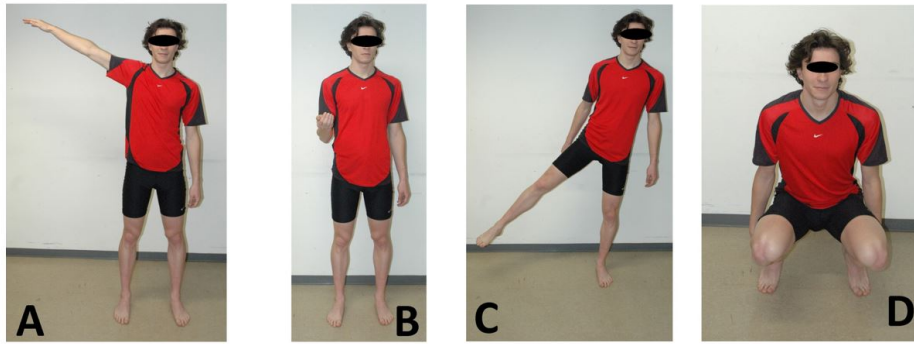


FIGURE 1.12: The movements performed in the study (Bonnechère et al., 2014) to compare the Microsoft Kinect camera with a marker-based motion capture system. (A) Shoulder abduction, (B) elbow flexion, (C) hip abduction, (D) knee flexion.

(Tanaka et al., 2018) examined the Microsoft Kinect v2 sensor and SDK's validity for measuring the lower limb joints kinematics against a marker-based motion capture system. Fifty-one healthy subjects walked while being captured by marker-less and marker-based motion capture systems simultaneously. The hip and knee joint kinematics were computed using both systems. The results indicated that the largest difference between the systems was less than or equal to 4.2° . Even though the results seemed encouraging, for measuring the difference, the bias between the system was removed by conducting the "parallel transformation" on the Kinect data. In a more recent study, (Wochatz et al., 2019) evaluated the validity and reliability of the Kinect v2 camera and the Microsoft Kinect SDK as a marker-less motion capture system against a marker-based motion capture system to measure hip and knee joint kinematics during rehabilitation exercises. Twenty-two healthy subjects did five repetitions of rehabilitation exercises, including squats, hip abduction, and lunge in two different sessions. The marker-less and marker-based systems' agreement was assessed using the Pearson correlation coefficient (r). The r ranged from 0.01 – 0.83, indicating weak to strong correlation. The standard error of measurement, representing the inter-session reliability, was, on average, 7.6° and 5.4° for marker-less and marker-based systems, respectively. Even though the marker-less system's reliability is comparable to the marker-based system, the system's validity is limited.

(Clark et al., 2019) authored a recent literature review on the uses, validity, and developments of the depth cameras, especially the Kinect camera. This study reviewed the validity of depth cameras for gait assessment of different motor tasks, including balance and functional tasks, posture, overground gait, and treadmill gait. For example, for the overground gait, it is reported that the Kinect's validity for most of the kinematic gait parameters was restricted ($r < 0.75$), and merely some spatiotemporal gait parameters such as step length and step width were valid ($r > 0.75$) (Springer and Yoge Seligmann, 2016). Also, based on several studies (Pfister et al., 2014; Xu et al., 2015), there are substantial errors in assessing kinematic gait parameters for the treadmill gait (RMS error $> 10^\circ$ or Bland-Altman limits of agreement $> 10^\circ$).

RGB cameras

The marker-less systems based on RGB cameras compute the gait parameters using two different approaches. DARI Motion's, The Captury's, and SIMI Motion's marker-less motion capture systems' work process is background subtraction, subject's silhouettes estimation, fitting a 3D body model to the silhouettes, and computing the gait parameters (refer to Figure 1.14). We refer to this approach as the **Image Analysis approach**. On the other hand, wrnchAI's and

Theai3D's marker-less motion capture systems work based on Artificial Intelligence (AI). We call this approach the **AI-based approach**.

Image analysis approach (Ceseracciu, Sawacha, and Cobelli, 2014) compared a marker-less motion capture system with a marker-based system in terms of the clinically relevant gait parameters. First, a 3D human body model of a subject was obtained using a laser scan. Then, eight cameras recorded the subject while walking. After background subtraction and silhouette extraction, the obtained 3D model was fitted to the extracted silhouettes to estimate the human poses at each video frame, and the gait parameters were calculated. The RMS differences between the marker-less and marker-based systems were 17.6° and 11.8° for the hip and knee extension-flexion angles, respectively. (Sandau et al., 2014) presented a similar study. However, apart from the differences in the incorporated methods for silhouette extraction and 3D model fitting, some technical aspects were different. First, the resolution of cameras increased. Second, the 3D human body model was acquired with the same setup, not a sperate 3D laser scanner. Third, the subject's wearing was a full-body snow leopard suit. The results showed that the RMS differences between the marker-less and marker-based motion capture systems were 2.6° and 3.5° for the hip and knee extension-flexion angles, respectively. There are notable differences between the results of these two studies. One of the underlying reasons might be the technical differences in terms of cameras and subject's wearing. The other reason might be the model generation process used in the second study that joint centers and body segment's coordinate frames were directly transferred from the marker-based system to the model.

DARI Motion's commercial marker-less motion capture system is called DARI Health. Querying the PubMed dataset with the keywords "DARI Motion" or "DARI Health" do not display any relevant results to marker-less motion capture systems. DARI Health utilizes the recorded videos by several RGB cameras (e.g., eight cameras) to quantify full-body human kinematics and kinetics. In March 2019, DARI Motion received the U.S. Foods and Drugs Administration (FDA) clearance⁴ for the DARI Health system. In the submitted document to the FDA, the results of a few studies were mentioned that assessed the performance characteristics of the DARI Health system. The studies found no statistically significant difference in clinically relevant joint kinematics ($p=0.33$) compared with a marker-based Vicon motion capture system. However, there was no information about the type of motion (e.g., walking or squat), statistical methods, and gait parameters. Though FDA cleared, DARI Health's number of cameras is relatively high, affecting the system's price. Figure 1.13 shows the required space for the DARI Health system setup, pointing out that a dedicated laboratory environment is required. Also, there is no published literature assessing the DARI Health system's accuracy or utilizing in clinical research.

Querying the PubMed dataset with the keyword "Captury" resulted in one article (Harsted et al., 2019) that has assessed the marker-less motion capture system's concurrent validity. (Harsted et al., 2019) explains that **The Captury** marker-less motion capture system uses a visual hull and a background subtraction method to estimate the subject's silhouette. Then, a subject-specific skeleton is fitted into the estimated silhouettes. Fourteen children (three to six years old) participated in this study performing a series of actions, including squats. Different variables were measured, including hip and knee extension-flexion. The mean difference, lower and upper limits of agreement in squats for knee extension-flexion were -11° , -25° , and 4° , for hip extension-flexion were -1° , -33° , and 32° , respectively. Though the values show that the Captury system's accuracy is not sufficient for clinical applications, it should be noted that

⁴510K Reference Number: K180880

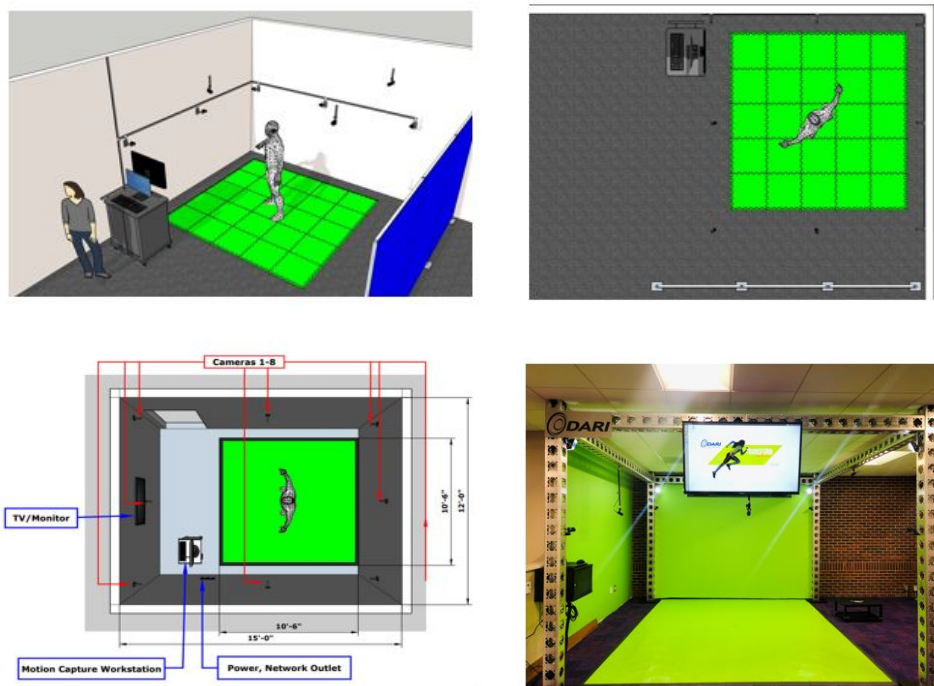


FIGURE 1.13: The hardware configuration and space requirement for a DARI Health marker-less motion capture system. Reproduced from DARI Motion.

the study population was only children. More comprehensive research should assess the concurrent validity of the system for different populations performing various motor tasks. The Captury and DARI Motion marker-less motion capture systems utilize the same technology to capture and estimate the human motion⁵. Thus, the Captury system's limitations in terms of the number of cameras and the required space are the same as the DARI Motions'.

SIMI Motion's products are marker-based and marker-less motion capture systems. As Figure 1.14 shows, the SIMI Motion's marker-less motion capture system performs the full-body motion capture by a series of steps. First, eight cameras record the subject whose wearings have good contrast to the background. Second, the subject's silhouettes are extracted. Third, a 3D human body model is fitted into the obtained silhouettes. Fourth, the 3D joints locations and joints angles are estimated. (Becker and Russ, 2015) assessed the accuracy of the SIMI Motion's marker-less motion capture system against a marker-based system. One subject performed twenty-two types of movement in all joints and planes. The standard deviations of angle difference for hip and knee extension-flexion were 17° and 4°. In another study, (Früh-schütz et al., 2017) evaluated the SIMI Motion's marker-less motion capture system's performance in sport. Various types of movements were recorded — specific joint movements (e.g., for the hip, knee, and ankle), complex movements (e.g., biking, running, and jumping), and highly dynamic movements (playing tennis). The standard deviation of difference for joint center locations over all the movements was 29 mm. Similar to some of the other commercial marker-less motion capture systems, the number of cameras (8 cameras) is relatively high. Also, more extensive research should validate the accuracy of the motion capture system for healthy and pathological subjects.

⁵In the FDA clearance document for the DARI Health system (510K Reference number: K180880), it is stated that "During movement, the DARI Health interfaces with the capture software (off-the-shelf, Captury Live) tracks all segments and joint centers as they move in all planes of motion."

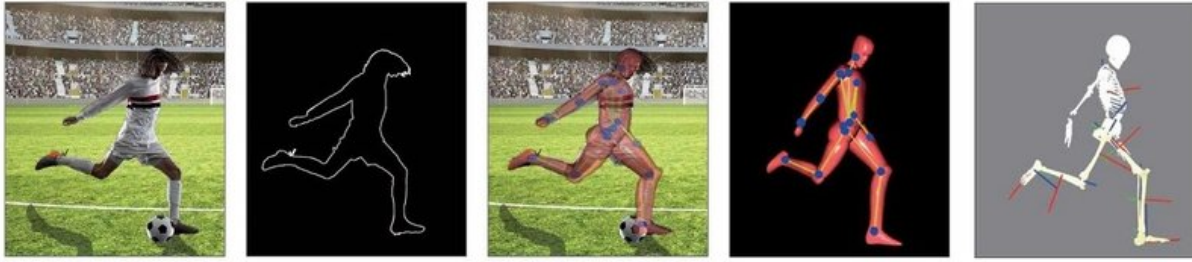


FIGURE 1.14: Working process with the SIMI Motion's marker-less motion capture system. From left to right: recording, silhouette extraction, 3D human body model fitting, and computation of 3D joint positions and joint angles. Reproduced from SIMI Motion.

The research studies based on the Image analysis approach of pose estimation, highlight several limitations. The primary limitations were a relatively high number of cameras and the requirement for the excellent contrast of the subjects wearing to the background. During the discussion on the commercial marker-less motion capture system, which implemented the Image analysis approach, these limitations were also observed.

AI-based approach Searching the PubMed dataset with the keywords "wrnchAI" and "Theai3D" do not display any relevant results to marker-less motion capture systems. The **wrnchAI** is a proprietary computer vision engine, powered by artificial intelligence to estimate human pose and motion — the wrnchAI works with any RGB or infrared cameras. Gupta (*Pose Detection comparison 2019*) recently evaluated the performance of the wrnchAI using a publicly available dataset (COCO dataset) containing 25,288 images in the validation and test data. The dataset contains only two-dimensional images and data, and the report determines the accuracy of the wrnchAI only at the two-dimensional level. There is no information about the performance of the system at the three-dimensional level. In terms of accuracy at the two-dimensional level, there are a few evaluation metrics for assessing the accuracy. In the following chapter, this issue will be discussed more comprehensively.

Theai3D The incorporated 3D human pose estimation method is based on artificial intelligence, primarily deep convolutional neural networks. (Kanko et al., 2020a) evaluated the performance of the Theai3D in measuring the spatiotemporal gait parameters. Theai3D used the recorded videos by eight synchronized RGB cameras to estimate the 3D human poses across the video frames. (Kanko et al., 2020a) claimed that the Bland-Altman method showed "no clinically meaningful differences" between the systems. In another study, (Kanko et al., 2020b) assessed the inter-trial and inter-session reliability of the gait kinematic data. The average inter-trial and inter-session errors for hip extension-flexion were 2.6° and 2.7°; for knee extension-flexion were 2.1° and 2.2°. Though there are a few reports that support the accuracy and reliability of the Theai3D marker-less motion capture system, the number of cameras is relatively high (8 cameras). Also, no peer-reviewed articles have evaluated the accuracy of the Theai3D for measuring the kinematic gait parameters, including clinically relevant joint angles, for healthy and pathological subjects.

The AI-based approach — based on artificial intelligence and especially deep learning — have attracted considerable attentions in recent years. For instance, (Clark et al., 2019), which authored a literature review on using depth cameras as marker-less motion capture systems, discussed the human pose estimation methods and stated the accuracy of human pose estimation methods, e.g., **OpenPose** (Cao et al., 2018) and **PoseNet**, for measuring kinematic joint

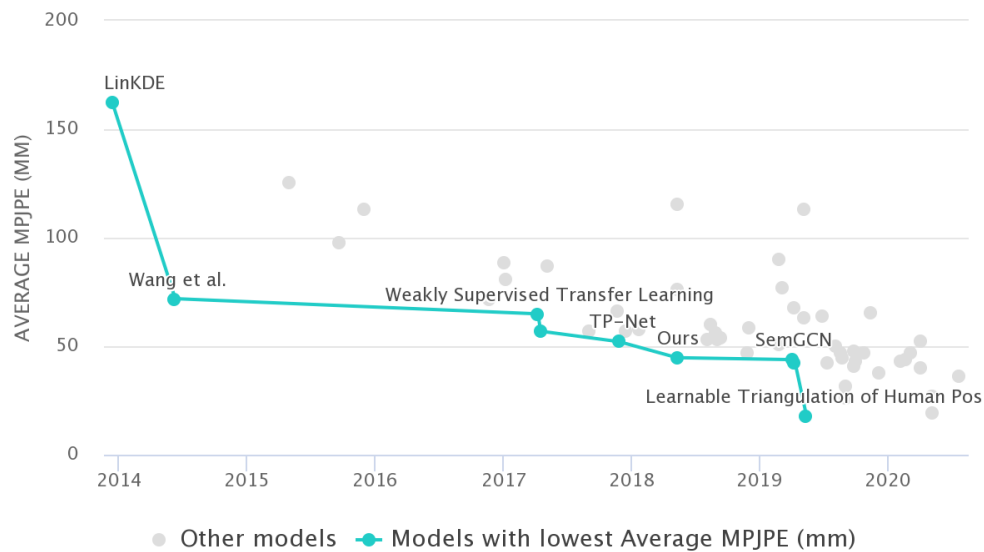


FIGURE 1.15: The AI-based 3D human pose estimation methods, evaluated on a publicly available dataset (Human3.6M) (*Human3.6M Benchmark*).

angles is unknown. The computer vision community have worked significantly on this research subject in recent years. Figure 1.15 shows the number of published research studies on AI-based 3D human pose estimation that evaluated their method only on a single publicly available dataset. In this figure, the implemented evaluation metric for assessing the 3D human pose estimation accuracy is MPJPE (Mean Per Joint Position Error). The lower MPJPE signifies the more accurate methods. This metric will be explained more comprehensively in the following chapter.

A summary of the characteristics of the marker-less motion capture systems based on RGB cameras is shown in Table 1.4. Except for the Theai3D system which uses a relatively high number of cameras, the performances of the other marker-less systems based on the AI-based approach in terms of spatiotemporal or kinematic gait parameters are not known. **The AI-based approach will be the focus of the next chapter.**

TABLE 1.4: Summary of the marker-less systems based on RGB cameras

		No.	Approach		Evaluation		
			AI	IA	Metric	Parameter	Results
Commercial	DARI Motion	8	□	■	Statistical difference	joint kinematics	$p = 0.33$
	The Capture	8	□	■	Bland-Altman LoA	hip & knee ext.-flex.	$-25^{\circ}, 4^{\circ}$ & $-33^{\circ}, 32^{\circ}$
	SIMI Motion	8	□	■	Standard deviation	hip & knee ext.-flex.	17° & 4°
	wrnchAI	1	■	□	-	-	-
	Theai3D	8	■	□	Bland-Altman LoA	spatiotemporal	"no difference" ¹
					Inter-trial error	hip & knee ext.-flex.	2.6° & 2.1°
Inter-session error					hip & knee ext.-flex.	2.7° & 2.2°	
Research	Sandau et al.	8	□	■	RMS error	hip & knee ext.-flex.	2.6° & 3.5°
	Ceseracciu et al.	8	□	■	RMS error	hip & knee ext.-flex.	17.6° & 11.8°
	OpenPose	1	■	□	-	-	-
	PoseNet	1	■	□	-	-	-

Abbreviations: **No.**: Number of cameras; **AI**: Artificial Intelligence; **IA**: Image Analysis; **LoA**: Limits of agreement; **RMS**: Root Mean Square; **ext.-flex.**: extension-flexion. ¹ "no clinically meaningful difference."

1.4 Conclusion

This chapter demonstrated the role and effectiveness of gait analysis for clinical conditions and fall risk assessment. Then, the procedure for measuring gait parameters using the current gold standard system and the reliability of gait parameters was studied. The most reliable kinematic gait parameters were hip, knee, ankle extension-flexion, pelvis ab-adduction, and rotation angles. The limitations of clinical gait laboratories were illustrated that were relevant mostly to marker-based motion capture systems. Thus, the alternatives for marker-based motion capture systems were comprehensively studied.

The IMU-based motion capture systems do not need a dedicated laboratory environment; But they have certain limitations in accuracy and costs. The marker-less motion capture systems based on depth cameras are easily affordable compared to marker-based motion capture systems; But the accuracy may not be adequate for clinical applications. The marker-less systems based on RGB cameras using the image analysis approach needs a high number of cameras, good contrast of subject's wearing with respect to the background, and a dedicated laboratory environment. Also, no study supports the adequacy of accuracy for clinical applications. The AI-based approach has attracted significant attention in recent years. Theai3D, founded in 2018, a commercial marker-less motion capture system based on the AI-based approach, has achieved promising results. However, this system requires a high number of cameras.

Numerous research studies focused on human pose estimation using RGB cameras and artificial intelligence techniques, achieving significant results in recent years. However, their performances in measuring clinically relevant gait parameters are unknown. We hypothesized that human pose estimation methods could provide a powerful tool for designing and developing an accurate, cost-effective, and easy-to-use **marker-less motion capture system** that can be considered an alternative for marker-based motion capture systems. The next chapter reviews the human pose estimation methods comprehensively.

Designing a system requires definite and quantified design criteria. Design criteria encompass the characteristics, including **costs**, **accuracy**, **required laboratory environment**, and **automaticity**:

1. The price of marker-based motion capture systems is comparatively high. The costs of required equipment for marker-less motion capture systems, including RGB cameras, installation equipment, connection equipment, calibration devices, synchronization devices, PC for data capture and processing, should be less than 5,000 €.
2. The marker-based systems are accurate enough for clinical applications. The marker-less motion capture system should be accurate and precise. The objective is to design a marker-less system as accurate and precise as the commercial marker-based motion capture systems to measure kinematic joint angles that are clinically relevant gait parameters. The quantified reliabilities of marker-based motion capture systems for measuring gait parameters are discussed in section 1.1.4. This section will assist as a guide for the accuracy of the marker-less motion capture system.
3. The need for a controlled laboratory environment and its associated costs (rent, electricity) limit the systems' utilization in other spaces. The marker-less motion capture system should be capable of being installed in diverse places, such as the corridor of a hospital or a physician's office. Also, there should be no necessity for any particular wearing.
4. As its name indicates, the designed **marker-less** system should not require any sorts of markers or sensors. The current marker-based motion capture systems require accurate

placement of markers on body anatomical landmarks by professional individuals. Consequently, labor cost is one of the factors that make the current gait analysis costly.

5. A gait study's needed time consists of patient preparatory time, testing time, and time for post-processing and analysis of data and report generation. The marker-less motion capture system drastically reduce the patient preparatory time. Also, the required time for post-processing and data analysis should be less than or equal to the marker-based system (30 minutes).
6. The marker-less motion capture system should automatically process the captured data.

Table 1.5 summarizes the design criteria for the marker-less motion capture system.

TABLE 1.5: The design criteria for the marker-less motion capture system

Priority	Characteristic	Description
very high	low-cost	$\leq 5,000$ €.
very high	accurate and precise	as accurate and precise as a commercial marker-based motion capture system
very high	no need for a controlled laboratory environment	no requirement for a specific wearing
high	Markerless	-
moderate	being Fast	not slower than a marker-based commercial motion capture system for data processing
moderate	automatic	-

Chapter 2

Human Pose Estimation

Human Pose Estimation (HPE) methods refer to computer vision techniques that can estimate human body keypoints positions such as joint centers or segment extremities positions. The objective of this chapter is to study the literature of **human pose estimation** that may help design a marker-less motion capture system for clinical applications.

Figure 2.1b shows an estimated two-dimensional human pose superimposed on the original RGB image, estimated using a convolutional neural network (ConvNet) (Iskakov et al., 2019). The ConvNet will be explained later in this chapter. The input to the ConvNet is the RGB image, and the output is a set of heatmaps, which indicates the probability of the presence of joint centers. Figure 2.1a shows the obtained heatmaps for the right ankle, knee, wrist, and shoulder. The maximum values of the obtained heatmaps are regarded as the estimated keypoints. In the following sections, we will clarify that ConvNets play a significant role in human pose estimation methods and are becoming the most used tool in this field of study.

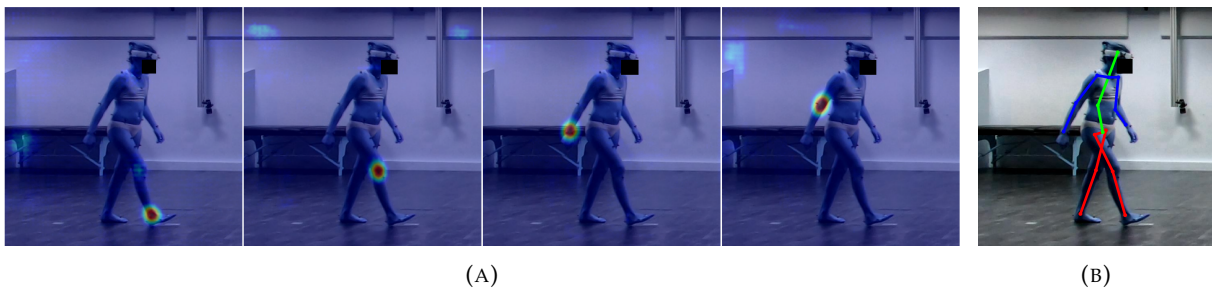


FIGURE 2.1: (A) The estimated heatmaps, superimposed on the RGB image, for right ankle, knee, wrist, and elbow. (B) The estimated two-dimensional human pose obtained using the maximum values of heatmaps.

2.1 Human pose datasets

There are many and different widely-used datasets for human pose estimation – e.g., CMU (*CMU Graphics Lab Motion Capture Database*), HumanEva (Sigal, Balan, and Black, 2010), LSP (Johnson and Everingham, 2011), MPII (Andriluka et al., 2014), COCO dataset (Lin et al., 2014), Human3.6M (Ionescu et al., 2014), CMU Panoptic Studio (Joo et al., 2015), MARCOI (Elhayek et al., 2016), SURREAL (Varol et al., 2017), 3DPW (Marcard et al., 2018). Datasets have different characteristics which make them appropriate for various applications. They consist of **still images** or **image sequences (videos)**, **single-view images** or **multi-view images (captured by several synchronized cameras)**, and specific or diverse **types of actions**. The datasets include annotations – e.g., the reference values for the body joint center locations – collected by different procedures. For example, the annotations of the MPII dataset are performed by in-house

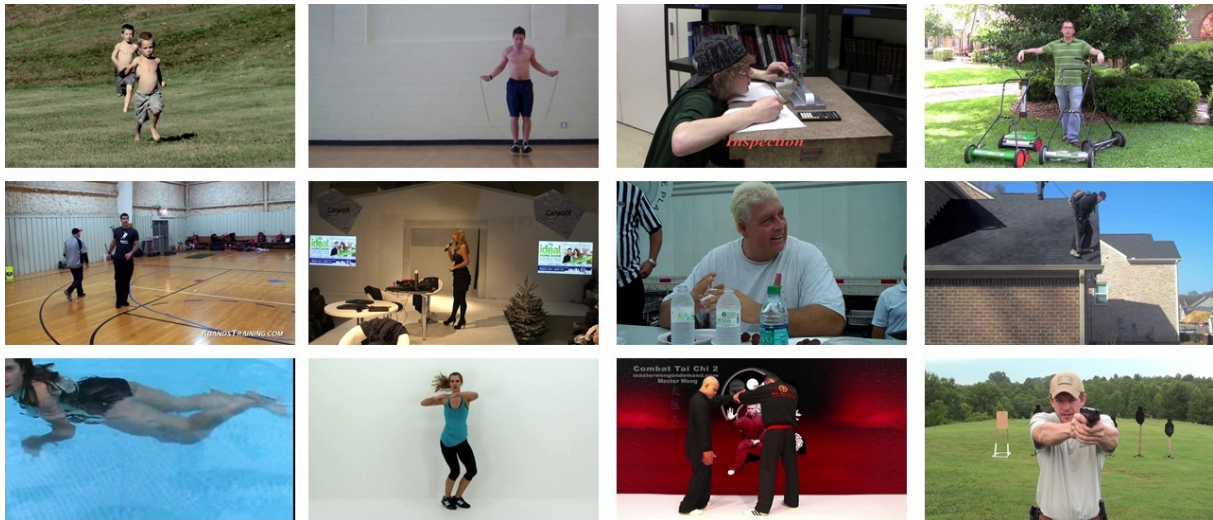


FIGURE 2.2: Sample images from the MPII dataset (Andriluka et al., 2014)

workers; The CMU, HumanEva, and Human3.6M datasets are the unique ones that utilized marker-based motion capture systems to provide the annotations. Herein, we focus on four datasets, MPII, CMU, HumanEva, and Human3.6M. MPII dataset is extensively used for evaluation of 2D HPE methods, and the remainings are proper for assessment of 3D HPE methods and are the only ones that used marker-based motion capture systems for annotations.

MPII dataset

MPII dataset (Andriluka et al., 2014) presents 40,522 images of people collected by querying YouTube. 491 activities were chosen, and for each activity, ten videos were selected which resulted in 3,913 videos. Then, several frames were picked from each video, with the point that the frames had to be 5 seconds apart. This resulted in 24,920 extracted frames and consequently images of 40,522 people. Several samples of the MPII dataset are illustrated in Figure 2.2.

CMU dataset

CMU dataset (*CMU Graphics Lab Motion Capture Database*) is one of the most extensive datasets. In this dataset, 2605 sequences captured from 109 subjects performing a diverse set of motions, are publicly available. However, for many sequences, only a low-resolution video captured by one camera is available (see Figure 2.3). On the other hand, the calibration information of cameras is not available, and the data captured by the marker-based motion capture system is not synchronized with the videos. All these factors make the performance evaluation difficult and affect the accuracy of the data.



FIGURE 2.3: Samples from the CMU dataset.

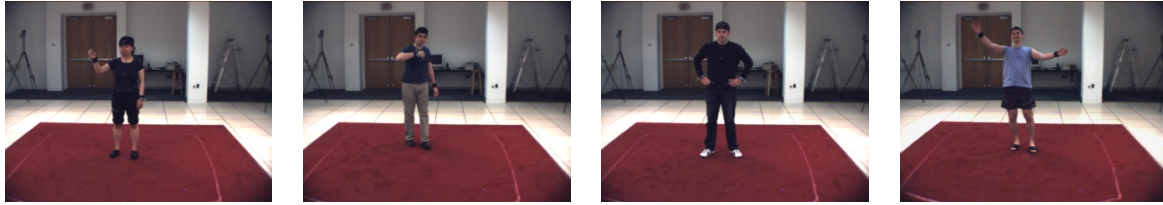


FIGURE 2.4: Samples from the HumanEva dataset (Sigal, Balan, and Black, 2009).

HumanEva dataset

HumanEva dataset (Sigal, Balan, and Black, 2009) was the first dataset that was suited for the accurate evaluation of 3D HPE methods. This dataset contains four subjects performing different motions captured by multiple synchronized cameras. In general, the elements, which are of importance to the datasets, are pose variability, appearance variability, and scale of the dataset. HumanEva dataset improved the pose variability by including different motions such as walking, jogging, throwing/catching, gesture, boxing, and combo. Subjects of the HumanEva dataset wore their regular clothing. Although it could decrease the accuracy of the motion capture system, this trade-off has been accepted to maintain realism as much as possible and increase the appearance variability. In terms of scales of the dataset, HumanEva contains around 80,000 frames or 56 sequences. Several samples of the HumanEva dataset are shown in Figure 2.4.

Human3.6M dataset

Human3.6M dataset (Ionescu et al., 2014) is the largest publicly available dataset that contains over 3.6 million diverse 3D human poses. The laboratory setup, used for data collection, consists of four RGB cameras and 10 Vicon Motion Capture cameras, synchronized in hardware level. The four calibrated cameras were placed at four corners of the rectangular capture area. 10 motion capture cameras helped retrieve the 3D position of reflective markers attached to the subject's regular clothing – even though this may reduce the accuracy; this trade-off was



FIGURE 2.5: Sample images from the Human3.6M dataset (Ionescu et al., 2014)

TABLE 2.1: Characteristics of the MPII, CMU, HumanEva, and Human3.6M

Dataset	MPII	CMU	HumanEva	Human3.6M
Number of subjects	40,522	109	4	7
Data type	single image	image sequence	image sequence	image sequence
Image type	single-view	single-view	multi-view	multi-view
Annotation	manual labelling	marker-based	marker-based	marker-based
Reference data	2D	3D	3D	3D
Wearing	regular	tight fitting	regular	regular
Camera calibration	No	No	Yes	Yes
Synchronized data	N/A	No	Yes	Yes
Pathological cases	No	No	No	No

made to maintain as much realism as possible. The position of reflective markers helped infer accurate 3D human body poses. Eleven subjects – seven subjects for training and validation, four subjects for testing – performed 15 different categories of motion, including directions, discussion, greeting, posing, purchases, taking photos, waiting, walking, walking a dog, walking pair, eating, phone talk, sitting, smoking, and sitting down. This dataset, in comparison to HumanEva, has improved the pose variability by increasing the classes of motions. Figure 2.5 shows several samples of the Human3.6M dataset.

Characteristics of the MPII, CMU, HumanEva, and Human3.6M datasets are summarized in Table 2.1. HumanEva and Human3.6M are the only datasets in which the reference data are synchronized with the image sequences and were obtained by marker-based motion capture systems. These two datasets have contributed significantly to the advancement of human pose estimation methods. However, they cannot be utilized for the specific application of gait analysis that accuracy is of essential importance. First, the subjects wore their regular clothing. The attachment of the markers to the regular clothing would reduce the accuracy of the motion capture acquisition. Second, the number of subjects is low. HumanEva and Human3.6M contain the motion of four and eleven subjects – the reference values for joint centers are only provided for seven out of eleven subjects. It limits the appearance variability. For example, all of Human3.6’s subjects were young adults. Third, since the objective of these datasets was different, they do not contain the motion of pathological cases. These elements highlight the need for a human pose dataset that is proper for the specific gait study application.

2.2 Evaluation metrics

There are different evaluation metrics used for measuring the performance of HPE methods in the literature. Herein, we review two of them that are broadly used to evaluate the performance of 2D and 3D human pose estimation methods.

- **Mean Per Joint Position Error (MPJPE)** (Ionescu et al., 2014) which is the mean Euclidean distance between the predicted joints and the corresponding reference values. The HumanEva dataset proposes an evaluation metric, referred to as **3D error**, that is the same as the MPJPE evaluation metric.
- **PCK** or **Percentage of Correct Keypoint** (Andriluka et al., 2014) – For a given estimated pose, a joint is considered correctly localized, if the Euclidean distance of predicted joint to the corresponding reference value, is less than a threshold. The threshold is a fraction of the bounding box size of the person. **PCKh** – It considers the bounding box size of

the head instead of the bounding box size of the person as the threshold. For instance, in **PCKh@0.5**, the threshold is 50% of the bounding box size of the head.

Although these metrics are conventional for comparing human pose estimation methods' performance, for the particular task of gait analysis, these metrics might not be of practical value. The kinematic parameters that are fundamental for gait study include, e.g., knee extension-flexion angle or hip lateral flexion angle. Hence, for gait analysis, a well-defined evaluation metric is required and should be determined. Also, the mean value is a limited indicator of the real likelihood, and other parameters such as the Bland-Altman lower and upper limits of agreement should be reported.

2.3 Evolution of human pose estimation

In this section, we demonstrate how convolutional neural networks have become a powerful tool in human pose estimation methods. To this end, we begin with the definitions of features and representation. We clarify the role of representation in the performance of the human pose estimation methods. Lastly, how the issue of determining the most competent representation has been addressed.

2.3.1 Features and representation

The data representation refers to any convention for the arrangement of information to enable information to be encoded and decoded (*Data Curation*). The data representation is of high importance. For instance, we may find it easier to explain a concept in our native language rather than a foreign language. These examples indicate the role of data representation. The same rule also applies to human pose estimation. Hence, the question is which representation should be utilized to represent an RGB image so that it helps achieve the most accurate pose estimates? To answer this question, we first review different data representations used in human pose estimation methods.

Features are pieces of information in the **representation** of data. Features is an all-inclusive term and may refer to several different information representations (Goodfellow, Bengio, and Courville, 2016). Herein, to adequately define the features of data representation for HPE, similar to the survey of (Gong et al., 2016), we classify the features into four different classes – **low-level**, **mid-level**, **high-level**, and **motion features**.

The **low-level features** are **information** such as silhouettes, contours, and edges that **represent** the shape, geometry, and appearance of the human body. Each feature has its specifications. For example, changes in textures or illuminations do not affect the silhouette; edges represent the different existing lines in the images; contours represent the outline of the body segments. Figure 2.6 shows a human body silhouette as of the low-level features. (Agarwal and Triggs, 2005), (Sapp, Toshev, and Taskar, 2010), and (Dimitrijevic, Lepetit, and Fua, 2006) utilized silhouettes, contours, and edges, respectively, for human pose estimation. However, since there was no extensive human pose dataset at the time, the performance of these methods was not evaluated on extensive pose datasets.

The **mid-level features** can be the **encoding** of **low-level features** that, in this way, we can also consider them as **feature descriptors**. Mid-level features can be a **combination** of low-level features and result in **reduced size** feature space. The shape context (Belongie, Malik, and Puzicha, 2001) is an example of mid-level features. To compute the shape context, a set of points should be sampled from the shape's contours. Then, the shape context represents



FIGURE 2.6: Low-level features: Silhouette feature extracted using the Local Graph Cut method in the MATLAB® Image Segmenter app.

the relative distribution of a set of points. Figure 2.7 shows an example of the shape context feature computation. Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Scale Invariant Feature Transform (SIFT) (Lowe, 1999) features are other examples of mid-level features.

(Yang and Ramanan, 2011) utilized HOG features for 2D human pose estimation. However, the performance was not evaluated on extensive human pose datasets. (Müller and Arens, 2010) utilized SIFT features for 2D human pose estimation and assessed the method's performance on the HumanEva dataset. The mean Euclidean distance for every joint or segment

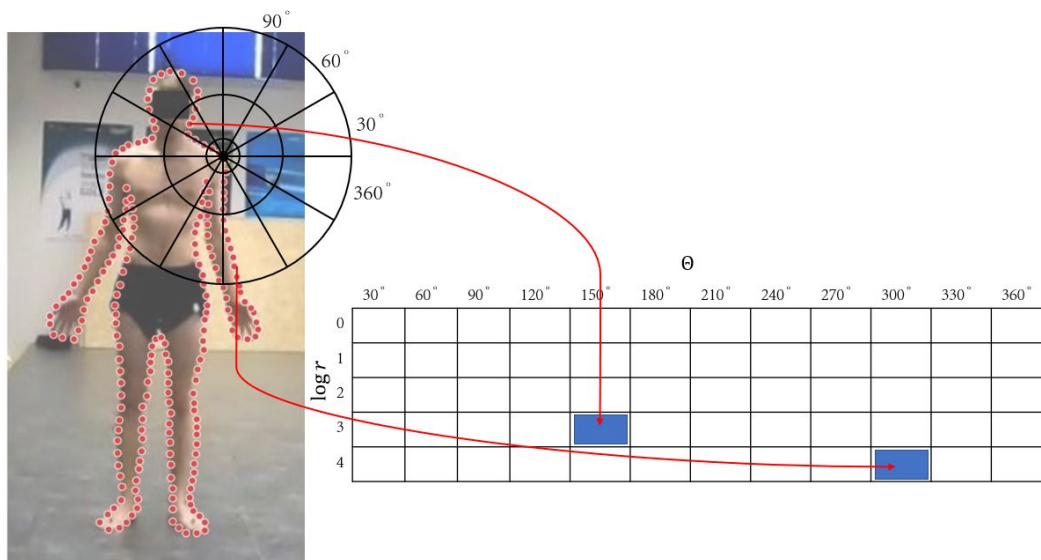
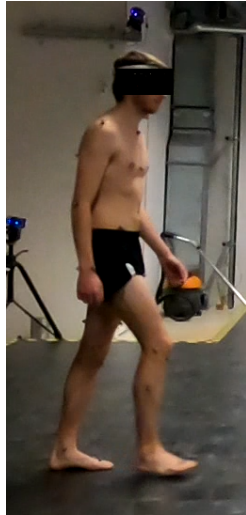


FIGURE 2.7: Mid-level features: Shape context computation. The shape context is encoded as a log-polar histogram of the coordinates of the points. In this instance, 5 bins for $\log r$ and 12 bins for θ are considered.



Example posebits

Right hand above the hips?	No
Right foot in front of the torso?	Yes
Left foot in front of the torso?	No
Left hand above the hips?	No
Right hand above the neck?	No
Left foot to the left of the hip?	No
Left hand to the left of the shoulder?	Yes
Right knee bent?	Yes
Right foot to the right of the hip?	No

FIGURE 2.8: High-level features: Posebits, as a high-level feature, consists of multiple yes/no questions and can be manually annotated.

centers varied from 7 to 72 pixels. (Amin et al., 2013) used shape context features and evaluated the performance on the HumanEva dataset. The results showed that the 3D error for the specific action of walking varied from 5 cm to 6 cm.

Several features can be considered to have high-level characteristics. For example, Posebits (Pons-Moll, Fleet, and Rosenhahn, 2014), as a geometry descriptor, represents the geometrical relationships between body parts. Figure 2.8 shows an example of Posebits. This study showed that by increasing the number of Posebits from zero to 15, the 3D error decreases from 11 cm to 7 cm. Another example is descriptive abstractions, realized in the (Oleinikov et al., 2014), that represent an image and target. For instance, the image description can be the state of clutter or occlusion (e.g., low, medium, and high); the target description can be the state of self-occlusion, size of the image, and pose (e.g., front-facing and regular). However, this method was not evaluated on extensive pose datasets.

Motion features This class of features is usually implemented to elevate the estimation performance using the temporal correspondence. For example, optical flow (Horn and Schunck, 1981), shown in Figure 2.9, is an estimate of two-dimensional image motion. (Lu and Jiang, 2013) utilized optical flow for human pose estimation. However, the study assessed the performance of the HPE method qualitatively, not quantitatively.

Thus far, we reviewed several data representations, from manually-designed low-level features to high-level and motion features. (Gong et al., 2016) and (Sarafianos et al., 2016) are two surveys that investigated the history of human pose estimation methods and provided numerous examples of employing manually-designed data representation. (Gong et al., 2016) states that besides features extraction algorithms, human pose estimation methods consist of different human body models and different methodologies. Therefore, the data representation is not the only parameter in a human pose estimation method. In other words, a combination of feature extraction algorithms, human body models, and methodologies determine the performance of HPE methods. Nevertheless, selecting a manually-designed features extraction algorithm for data representation that would achieve better performance in combination with human body models or methodologies remains an issue. In the following, we study how ConvNets-based human pose estimation methods attempted to address this issue.

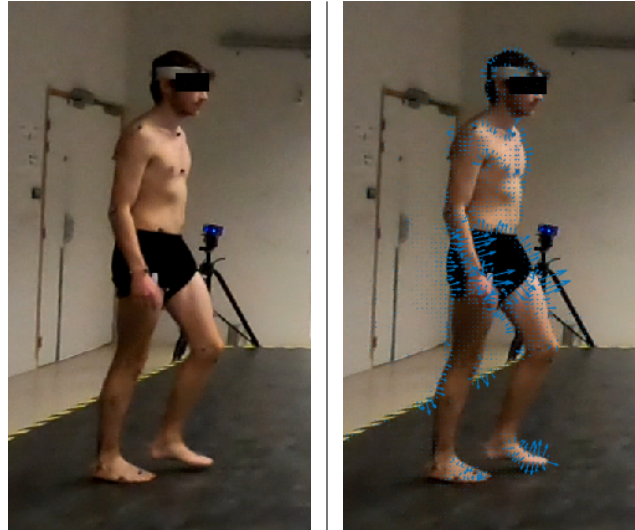


FIGURE 2.9: Motion features: Optical flow estimated using the Horn-Schunck method (Horn and Schunck, 1981).

2.3.2 Representation learning

Artificial Intelligence (AI) is an ever-growing field of study. This field was created in pursuing the dream of intelligent machines that can think. **Machine learning** is the ability of AI systems to acquire their knowledge by extracting patterns from raw data. For example, nowadays, machine learning algorithms can efficiently perform the separation of legitimate emails from spam emails (Goodfellow, Bengio, and Courville, 2016). Several applications of machine learning in radiology are image segmentation, image registration, computer-aided detection and diagnosis, and brain activity analysis (Wang and Summers, 2012).

Representation learning¹ is an approach in which machine learning is used to find the representation besides the mapping to output. **Learned representation usually achieve better performance than manually-designed representation.** **Deep learning** resolves the problem of representation. Deep learning is an approach that learns to express more abstract representations in terms of more simple representations (Goodfellow, Bengio, and Courville, 2016). Since 1980s (Fukushima and Miyake, 1982), deep learning popularity, performance, and usefulness have been and are still increasing over time. Two main reasons have significantly contributed to this achievement, **the availability of a massive amount of training data**, and **the availability of its required computer infrastructure**. Today, we see the application of deep learning in numerous fields, from speech recognition to image segmentation and human pose estimation. **Convolutional neural networks** are one of the networks which contributed significantly to the advancement of deep learning (Goodfellow, Bengio, and Courville, 2016).

2.3.3 Convolutional neural networks

ConvNets² are a specialized kind of neural network that employs a mathematical operation called convolution. The convolution of two real-valued functions can be written as below:

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t - \tau)d\tau \quad (2.1)$$

¹For further information, please refer to (Bengio, Courville, and Vincent, 2013)

²For more comprehensive information, please refer to (Gu et al., 2018)

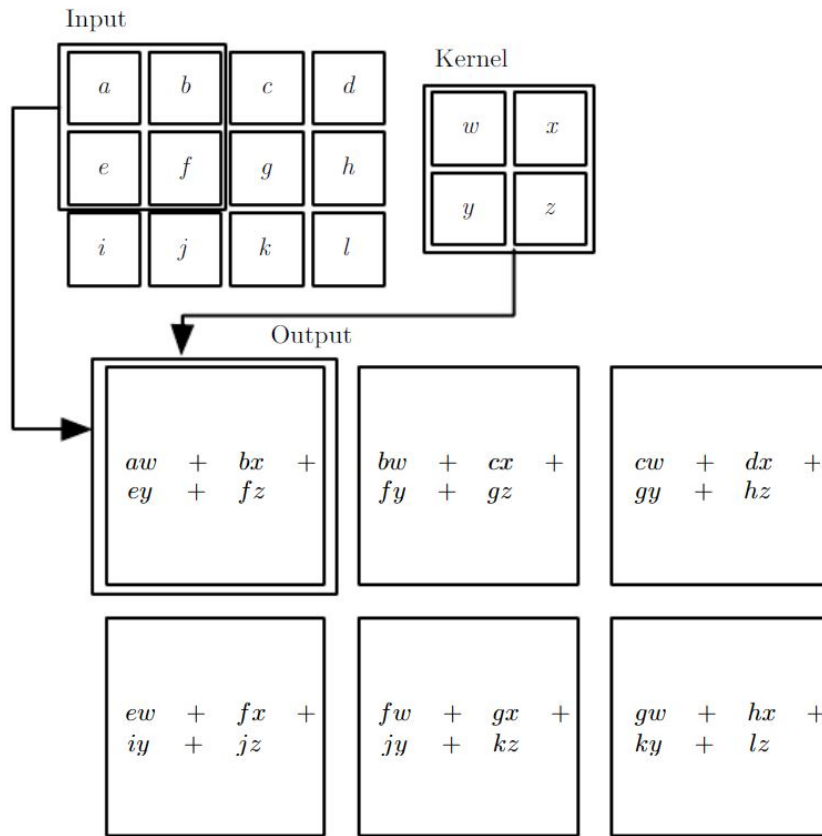


FIGURE 2.10: An example of convolution. Herein, the size of the output 2×3 is different from the input's 3×4 . (Goodfellow, Bengio, and Courville, 2016)

In **convolutional terminology**, the first function is referred to as the **input** (I) and the second one as the **kernel** (K). Also, the output ($I * K$) is referred to as the feature map. Since data on a computer is discretized, the discrete convolution is needed. In machine learning applications – e.g., computer vision – the input is usually a **multidimensional array**. Therefore, the discrete 2D convolution is required that can be written as below:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n) \quad (2.2)$$

where (i, j) and (m, n) are indices of the multidimensional arrays.

Many neural network libraries use a similar function called the cross-correlation; however, both of them are called convolution:

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n) \quad (2.3)$$

Figure 2.10 shows an example of the so-called convolution.

Pooling

A general layer of a convolutional network consists of three stages, convolution, activation functions like rectified linear activation function, and pooling function.

A pooling function substitutes a rectangular neighborhood with its summary statistic. There are different pooling operations like max pooling, average pooling, and weighted average pooling. Pooling is designed to help make the representation invariant to the local translations of input. Figure 2.11 shows an example of max pooling and average pooling operations.

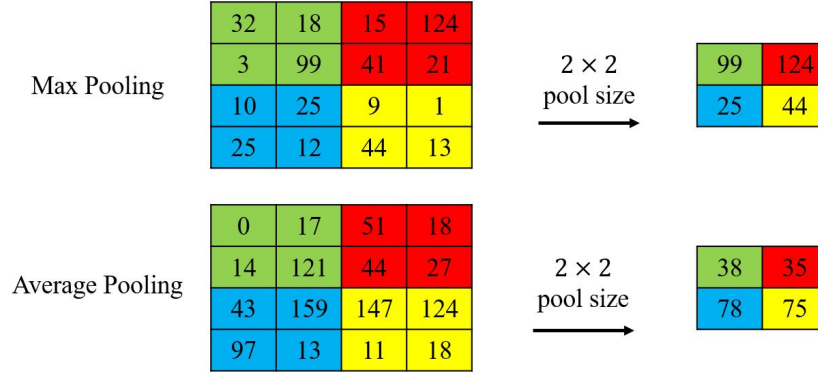


FIGURE 2.11: An illustration of max pooling and average pooling operations.

As an example of implementation of convolution and pooling operations in deep learning, Figure 2.12 illustrates the convolutional network architecture of a 2D human pose estimation method (Wei et al., 2016). As shown, the structure consists of multiple convolutions and pooling operations sequentially stacked across six different stages. The 368×368 pixels input images are passed to the network to regress the heatmaps of P parts (plus one for background).

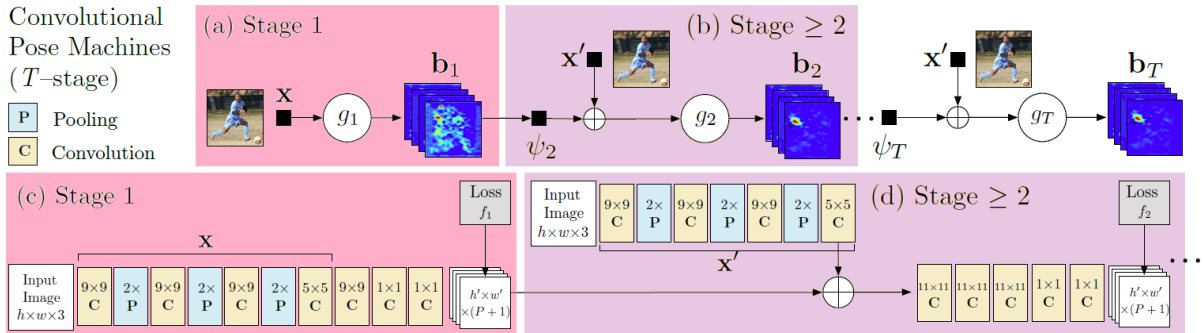


FIGURE 2.12: The convolutional network architecture of a state-of-the-art 2D human pose estimation method (Wei et al., 2016)

2.4 Human body model

Human body kinematic structure information, shape information, and appearance information are the main components of a human body model. These models help incorporate a priori knowledge – e.g., kinematic and dynamic constraints like motion limits of joints or motion patterns for specific activities – into the human pose estimation methods. The kinematic model could be utilized to assemble the detected body parts or joints, or these models with a pose could be projected to the image plane to verify the projected pose using the image observations (Gong et al., 2016). Different types of human body models are shown in table 2.13. In the following, we review several examples of these models that are extensively utilized in the literature of human pose estimation.

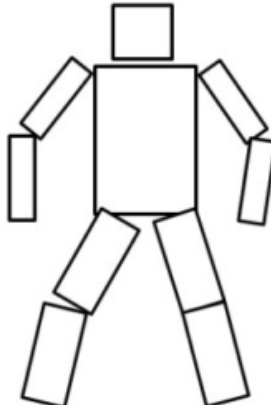
Kinematic model	Planar model	Volumetric model
Kinematic models show the skeletal structure of the human body.	These models, moreover to the information about the body parts and connecting relations, contains information about the appearance of the human body.	This kind of models realistically represent the 3D body shapes. Geometric shapes poses and meshes are two samples of the volumetric models.
		
(a)	(b)	(c)

FIGURE 2.13: Different types of human body models. (a) Kinematic model; (b) Cardboard model; (c) Volumetric model (Gong et al., 2016)

Kinematic model

A popular example of kinematic model is Pictorial Structure Model (PSM). PSMs provide a general framework to represent an object by a set of parts that some of them are connected (Felzenszwalb and Huttenlocher, 2005). Thus, we can consider it as an undirected graph $G = (V, E)$ where V represents the parts, and E represents the edges connecting some of the parts. PSM could be implemented either in 2D HPE methods (Johnson and Everingham, 2010; Belagiannis et al., 2014) or in 3D HPE methods (Qiu et al., 2019).

Planar model

The Contour Person (Freifeld et al., 2010) is an example of planar human body models. This 2D model is learned from a projected 3D model that captures the appearance and pose variations. The cardboard model (Ju, Black, and Yacoob, 1996) is another example that consists of a set of connected body parts represented by rectangular shapes.

Volumetric model

SCAPE (Shape Completion and Animation for PEople) (Anguelov et al., 2005) is a widely-used volumetric model that contains shape and pose information. This model is learned from two datasets; one includes scans of 70 poses of a single person, and the other one includes scans of 37 subjects in a similar pose. Skinned Multi-Person Linear (SMPL) model (Loper et al., 2015) is another volumetric human body model that can represent a broad range of human body shapes and poses. This model is learned from 1786 scans of different subjects in various poses. SMPL model $M(\vec{\beta}, \vec{\theta}) : \mathbb{R}^{|\vec{\beta}| \times |\vec{\theta}|} \mapsto \mathbb{R}^{3N}$ maps the shape and pose parameters to $N = 6980$ vertices.

2.5 Human pose estimation: State-of-the-arts

We can classify the human pose estimation methods into two-dimensional (2D) and three-dimensional (3D) methods. As may be evident, 2D HPE methods determine the pixel locations of the human body joint centers, while 3D HPE methods determine the 3D position of joint centers with respect to a coordinate system. Even though numerous 3D HPE methods rely on the 2D HPE methods, several 3D methods infer the final human body pose without any intermediate 2D pose. Also, publicly available datasets and the evaluation metrics are different for the 2D and 3D methods. In the following, after studying the human body model, we will present the state-of-the-art 2D and 3D HPE methods.

2.5.1 2D human pose estimation methods

In two different periods, we attempted to find the recent advancements in 2D HPE methods. First, in the second semester of 2017, we queried Scopus Elsevier with the keywords "pose estimation" AND "human". Second, in the second semester of 2019, we queried Google Scholar with the keyword "pose estimation." Then, documents which were related to hand pose estimation, head pose estimation, multi-person pose estimation, 3D human pose estimation, and depth cameras were filtered out. Also, since there were numerous articles focused on 2D HPE methods, for a better analysis, we concentrated on the papers that utilized the MPII dataset for performance evaluation. Finally, we reached to 30 articles.

It is worth noting that studying 2D HPE methods is of value because they form the basis for numerous 3D HPE methods. Therefore, there is a direct relation between the performance of 2D and 3D HPE methods. The performance of HPE methods in terms of accuracy could be considered as the most critical factor for their evaluation. Therefore, we first attempted to gain an overview of the performance of 2D HPE methods. Table 2.2 shows the achieved results of 2D HPE methods on the MPII dataset using the PCKh@0.5 evaluation metric.

Different human body joint centers cannot be estimated with the same level of accuracy. Figure 2.14 confirms this statement. The easiest joint centers to detect is head and the most difficult one is ankle. In fact, what distinguishes the performance of 2D HPE methods is their ability to estimate the most challenging joint centers – e.g., wrists, knees, and ankles. It should not be surprising because wrists, elbows, knees, and ankles are the joint centers that can significantly affect the human pose variability. Therefore, learning-based methods will face difficulties in learning the high pose variability across the dataset.

The interesting point is that all the papers, presented in Table 2.2, employ ConvNets to learn and estimate the various human poses. The differences among the methods are network architecture, training strategy, input, and output resolution. The network architecture could be considered as the beating heart of the 2D human pose estimation methods. The arrangement of different layers – e.g., convolution and pooling layers – and the connections among the layers that designate the flow of information across separate layers determine the network architecture. Each 2D human pose estimation method has its unique network architecture. However, several network architectures, such as **GoogleNet** (Szegedy et al., 2015), **VGGNet** (Simonyan and Zisserman, 2014), **Hourglass network** (Newell, Yang, and Deng, 2016), **ResNet** (He et al., 2016), and **Convolutional Pose Machines** (Wei et al., 2016), are utilized in various human pose estimation methods. The network architecture of **Convolutional Pose Machines (CPMs)** is already presented in Figure 2.12. Though, at that time, we did not describe it as conventional network architecture. In the following, we will review the **VGGNet**, **ResNet**, and **Hourglass network** architecture shortly. As Table 2.3 shows, these network architectures are utilized in

TABLE 2.2: 2D HPE methods, and their achieved results (*percentage*) on the MPII dataset using the PCKh@0.5 evaluation metric.

Paper	head	shoulder	elbow	wrist	hip	knee	ankle	total
(Bulat and Tzimiropoulos, 2017)	94.7	89.6	78.8	71.5	79.1	70.5	64	78.1
(Hu and Ramanan, 2016)	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
(Pishchulin et al., 2016)	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
(Rafi and Gall, 2016)	97.2	93.9	86.4	81.3	86.8	80.6	73.4	83.6
(Liang et al., 2018)	90.4	91.7	96.4	84.0	82.5	76.5	71.3	84.1
(Belagiannis and Zisserman, 2017)	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
(Bem et al., 2018)	97.7	95.0	88.1	83.4	87.9	82.1	78.8	88.1
(Insafutdinov et al., 2016)	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
(Wei et al., 2016)	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
(Kawana et al., 2018)	97.7	95.8	90.1	85.6	88.8	84.8	81.7	88.5
(Bi and Zou, 2019)	96.5	95.3	89.9	84.8	88.3	84.7	81.0	88.8
(Bulat and Tzimiropoulos, 2016)	97.9	95.1	89.9	85.3	89.4	85.7	81.9	89.7
(Zhang, Zhu, and Ye, 2019a)	98.3	96.0	90.9	86.7	90.2	87.0	83.6	90.8
(Newell, Yang, and Deng, 2016)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
(Zhang, Zhu, and Ye, 2019b)	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
(Ning, Zhang, and He, 2018)	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
(Tang et al., 2018)	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
(Chu et al., 2017)	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
(Liu et al., 2018)	98.4	96.4	92.0	87.8	90.7	88.3	85.3	91.6
(Yang et al., 2017)	98.4	96.5	91.9	88.2	91.1	88.6	85.3	91.8
(Chou, Chien, and Chen, 2018)	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
(Zhu et al., 2019)	98.1	96.7	92.5	88.4	90.8	88.8	95.3	91.8
(Chen et al., 2017)	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
(Ke et al., 2018)	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
(Sun et al., 2019)	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
(Tang, Yu, and Wu, 2018)	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
(Nie et al., 2018)	98.6	96.9	93.0	89.1	91.7	89.0	86.2	92.4
(Zhang et al., 2019a)	98.6	97.0	92.8	88.8	91.7	89.8	86.6	92.5
(Wang et al., 2019d)	98.5	97.1	92.7	88.9	91.6	89.6	86.6	92.5
(Tang and Wu, 2019)	98.7	97.1	93.1	89.4	91.9	90.1	86.7	92.7

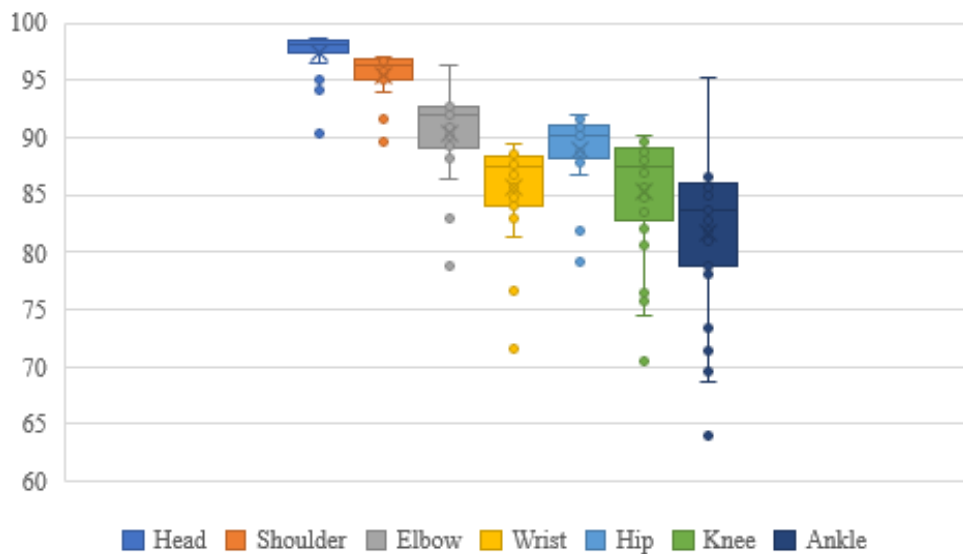


FIGURE 2.14: Performance of 2D HPE methods, presented in Table 2.2, for estimation of different body joint centers, evaluated using the PCKh metric.

different human pose estimation methods.

Visual Geometry Group (VGG) from the University of Oxford proposed **VGGNet** (Simonyan and Zisserman, 2014) for the ImageNet challenge 2014. The contribution of this network was that, despite previously proposed network such as **AlexNet** (Krizhevsky, Sutskever, and Hinton, 2012), it used a relatively small (3×3) convolution kernel and increased the depth to 16-19 layers. Figure 2.15 shows the network architecture of VGG-19 that consists of 19 layers and 144 million parameters.

Residual Networks (ResNets) (He et al., 2016) were proposed for the ImageNet challenge 2015. In ResNets, in addition to the convolutional and pooling layers, there are shortcut connections that perform identity mapping. Thanks to their particular structure, these networks, in comparison to the VGGNets, are easier to optimize, and their depth can be increased up to 152 layers. These characteristics result in improved performance of the ConvNet. Figure 2.15 shows the network architecture of ResNet-34 consisting of 34 layers.

Even though **VGGNets** and **ResNets** were designed for the classification task, the **Hourglass network** (Newell, Yang, and Deng, 2016) was designed to address the task of human pose estimation. That is why it became the basis for various HPE methods. The motivation behind the design of the Hourglass network is the requirement to capture information at different scales. In Hourglass networks, convolutional and max-pooling layers process feature down to a low resolution and then upsample and combine features.

In 2016, the Hourglass network was proposed, and to the best of our knowledge, it was the first 2D HPE method that its performance evaluated using the PCKh metric scored over 90% (90.9%). From 2016 to 2019, various network architectures were proposed. (Tang and Wu, 2019) achieved the best results (92.7%) on the MPII dataset. However, the results in comparison to Hourglass Network improved **only** by 1.8%. On the other hand, other factors affect the performance of the ConvNet-based human pose estimation methods, including training strategy, pre-processing of the dataset, and data augmentation. Therefore, the comparison of the human pose estimation methods and, in particular, their network architectures seems to be a challenging task. Nevertheless, we can admit that network architectures such as **VGGNets**, **ResNets**, **Hourglass network**, and **Convolutional Pose Machines (CPMs)** present an outstanding foundation for the task of human pose estimation. In the following subsection, we will see that these networks are extensively utilized for the task of 3D human pose estimation.

TABLE 2.3: The human pose estimation methods which utilized **VGGNet**, **ResNet**, and **Hourglass networks** in their network structure.

Network architecture	No.	References
VGGNet	2	(Nie et al., 2018; Bulat and Tzimiropoulos, 2016)
ResNet	4	(Insafutdinov et al., 2016; Liang et al., 2018; Bulat and Tzimiropoulos, 2016; Zhang et al., 2019a)
Hourglass network	14	(Bulat and Tzimiropoulos, 2017; Chu et al., 2017; Yang et al., 2017; Wang et al., 2019d; Zhang et al., 2019a; Tang, Yu, and Wu, 2018; Tang et al., 2018; Zhang, Zhu, and Ye, 2019b; Ke et al., 2018; Tang and Wu, 2019; Zhu et al., 2019; Ning, Zhang, and He, 2018; Chou, Chien, and Chen, 2018; Nie et al., 2018)

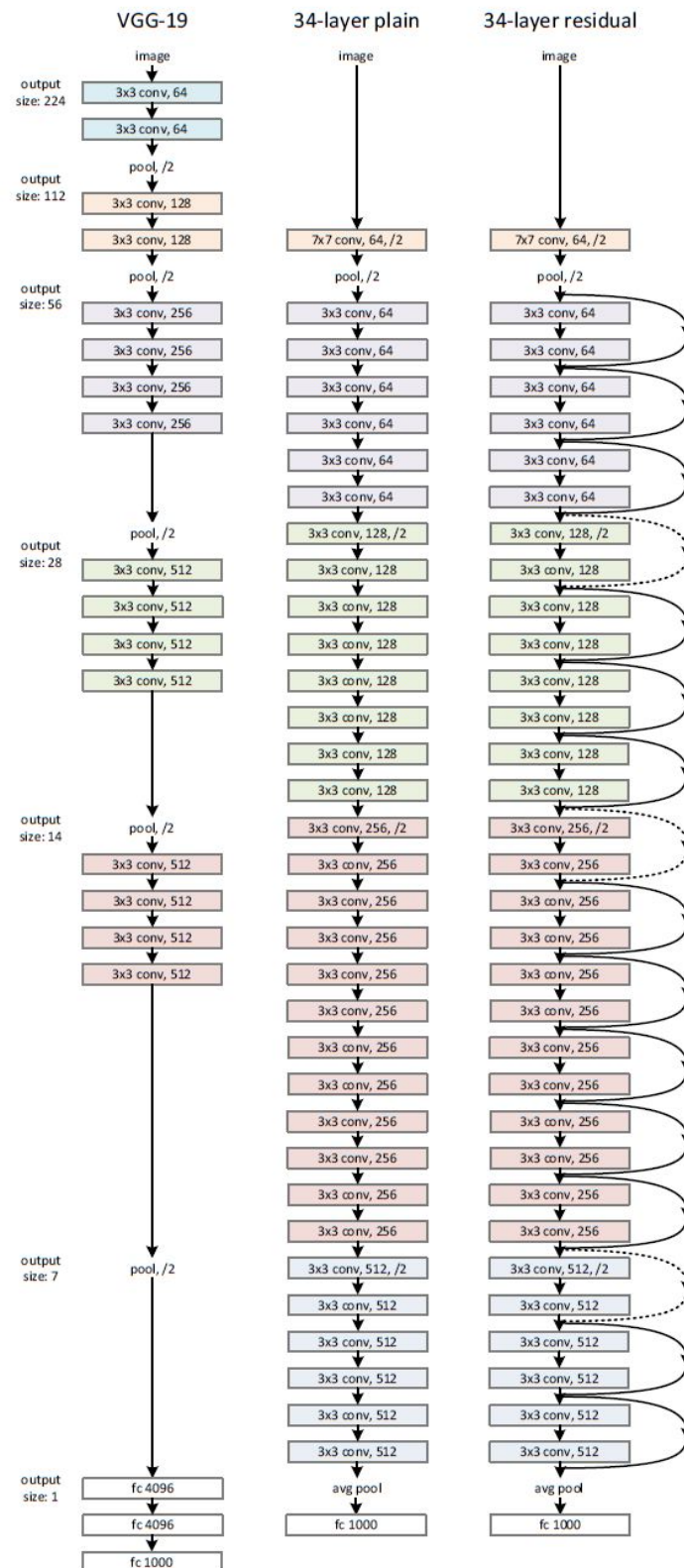


FIGURE 2.15: **Left:** the network architecture of the VGG-16. **Middle:** a plain network with 34 layers. **Right:** the network architecture of ResNet-34 with 34 layers. The dotted skip connections increase dimensions. (He et al., 2016)

2.5.2 3D human pose estimation methods

To study the recent articles in the field of 3D human pose estimation, in the second semester of 2019, we queried google scholar by the keyword "pose estimation" to find the latest studies – from 2018. Then, the articles related to 2D pose estimation, hand pose estimation, head pose estimation, multi-person pose estimation, object pose estimation, depth cameras, pose transfer, clothed human body shape reconstruction were filtered out. Finally, 41 articles met our criteria. To differentiate the 3D HPE methods, besides human body model, the input type is considered as another element of classification. The four RGB image **input types** are, **a still monocular image**, **a sequence of monocular images**, **a still multi-view image**, and **a sequence of multi-view images**.

Herein, we analyze the methods according to their performance on publicly available datasets. 37 out of 41 articles utilized the Human3.6M dataset to evaluate the performance of their proposed methods. This opportunity provides the possibility to perceive how the main elements of 3D HPE methods – e.g., input type, human body model, methodologies – can contribute to their performance. Even though the Human3.6M was used by most of the articles, the evaluation protocols were different. Therefore, we set a definition of protocols to be able to compare the methods:

- Protocol 1: Five subjects are used for training. Two subjects are used for evaluation. The MPJPE metric is calculated without any further post-processing on estimated 3D poses. The number of extracted frames could be different – e.g., every 5th frames or every 64th frames.
- Protocol 2: Five subjects are used for training. Two subjects are used for evaluation. The MPJPE is calculated after post-processing – it can be either the alignment of the root joint or the Procrustes method. The number of extracted frames could be different.

The defined protocol might not provide an accurate comparison; however, this helps examine all the 37 methods. Table 2.4 shows the evaluation results of articles on the Human3.6M dataset. Also, the use of a human body model and type of input to the 3D HPE methods are shown. This table has two sections – top and bottom – which is separated by a horizontal line. In the top section, results are sorted based on the protocol 1, and in the bottom section, based on the protocol 2. In these protocols, since the reference data for joint centers were only available for seven subjects, these protocols split the dataset into five and two subjects for training and test sets. However, this protocol may not be applicable as a test set for clinical studies because of a limited number of subjects in the test set.

(Iskakov et al., 2019) and (Qiu et al., 2019) achieved the-state-of-the-art results (MPJPE = 17.7 mm, 26.2 mm) on the Human3.6M dataset. There are some similarities between these two methods. In both methods, the input type is **a still multi-view RGB image**; The incorporated network architecture is **ResNet-152**; the **camera calibration parameters** are used to retrieve the 3D position of joint centers.

As shown in Table 2.4, most of articles (62.2%) attempted to address the 3D human pose estimation using a still monocular RGB image. Normally, recovering 3D human poses using monocular images is a challenging task in computer vision due to issues like occlusion and ambiguity between 2D and 3D poses.

16.2% of articles used a sequence of monocular RGB images to estimate 3D poses, aiming to use temporal information to better accuracy. Also, 16.2% of articles tried to increase the number of views to overcome the difficulties of like occlusion. Generally, these methods are

more accurate than the methods based on still monocular images – they are in the upper half of the Table 2.4. However, sometimes the accuracy of the methods based input type of a still monocular image is more accurate than a still multi-view image or a sequence of still monocular images. For instance, the accuracy of (Wang et al., 2019c), and (Biswas et al., 2019) is better than (Chen et al., 2019b) and (Kocabas, Karagoz, and Akbas, 2019).

Yet, we are sure that a multi-view image or a sequence of still images affords much more information than a still monocular image. Hence, **we believe that the performance of 3D human pose estimation methods depends on "what" sorts of information and "how" they are applied to estimate the human body pose.** Herein, "what" represents sorts of information such as input type, body model, camera calibration parameters, and temporal information; "how" describes the network architectures, aggregation method of extracted features, and even training strategy and its hyper parameters.

The methods based on a sequence of multi-view images may confirm this hypothesis. Even though these methods (only 5.4% of papers) utilize an abundant source of information, their performance is even less than several methods based on still monocular images.

A sort of information that was successfully exploited to improve the accuracy of 3D HPE methods is camera calibration parameters. These parameters can tremendously contribute to the performance of 3D HPE methods. (Qiu et al., 2019; Gundavarapu et al., 2019; Iskakov et al., 2019), are three examples that outperform the other methods and used camera calibration parameters in their 3D HPE algorithms.

Regarding human body models, as a sort of information that can be exploited, the methodologies which employed a model did not necessarily achieve the best performance. However, this point does not mean that a human body model cannot improve the accuracy; but shows that the methods of employment of the human body model were not practical. It is worth recalling that in the case of the SMPL human body model, the output is not only a 3D pose but also a 3D mesh; therefore, the 3D pose is just one of the factors that determine the performance.

If we focus on the used convolutional network architectures as a part of "how" available information is processed and interpreted to estimate 3D poses, we see that the conventional structures in 2D HPE methods are prevailing in 3D HPE methods. However, it should be noted that comparing "how" information is used is a demanding task. For example, consider two articles, (Iskakov et al., 2019) and (Gundavarapu et al., 2019). Both studies use still multi-view images to estimate 3D poses. Both of them estimate the 2D poses and the corresponding estimation confidence using the ResNet network architecture to furtherly use for 3D pose estimation. Even though both of them are among the most accurate methods; there is a significant difference between their achieved results. It indicates that differences in the details of the methods result in significant differences in MPJPE – 23.8 mm versus 32.7 mm. Herein, the reported MPJPE value for (Iskakov et al., 2019) was different because this value is obtained by running the provided `code` on the whole Human3.6M evaluation dataset, whereas the reported value in the Table 2.4 is taken from the paper.

TABLE 2.4: Evaluation of different studies on Human3.6M dataset. The numbers are taken from the respective papers. **Input 1**: Input type – a still monocular RGB image; **Input 2**: Input type – a sequence of still monocular RGB image; **Input 3**: Input type – a still multi-view RGB image; **Input 4**: Input type – a sequence of multi-view RGB image; **Model**: Human body model; **MPJPE 1**: Mean per joint position error (*mm*) – protocol 1; **MPJPE 2**: Mean per joint position error (*mm*) – protocol 2. The upper part of the table’s methods are sorted according to protocol 1; the lower part is sorted according to protocol 2.

Paper	Input 1	Input 2	Input 3	Input 4	Model	MPJPE 1	MPJPE 2
(Qiu et al., 2019)	□	□	■	□	■	26.2	
(Gundavarapu et al., 2019)	□	□	■	□	□	32.7	
(Wang et al., 2019c)	■	□	□	□	□	37.6	
(Biswas et al., 2019)	■	□	□	□	■	42.8	
(Chen et al., 2019b)	□	□	■	□	□	46.3	41.6
(Pavlo et al., 2019)	□	■	□	□	□	46.8	36.5
(Kocabas, Karagoz, and Akbas, 2019)	□	□	■	□	□	51.8	45.0
(Rayat Imtiaz Hossain and Little, 2018)	□	■	□	□	□	51.9	42.0
(Liu et al., 2019a)	■	□	□	□	□	52.4	38.4
(Tome et al., 2018)	□	□	■	□	■	52.8	44.6
(Liu, Akhtar, and Mian, 2019)	□	■	□	□	■	54.0	52.3
(Sárándi et al., 2018)	■	□	□	□	□	54.2	
(Núñez et al., 2019)	□	□	□	■	□	54.2	
(Pavlakos, Zhou, and Daniilidis, 2018)	■	□	□	□	□	56.2	41.8
(Wei et al., 2019)	■	□	□	□	□	56.6	42.8
(Sharma et al., 2019)	■	□	□	□	□	58.0	40.9
(Huang et al., 2017)	□	□	□	■	■	58.2	
(Guler and Kokkinos, 2019)	■	□	□	□	■	60.3	46.5
(Liu et al., 2019b)	■	□	□	□	□	61.1	
(Wang et al., 2019b)	□	■	□	□	□	63.7	
(Sun et al., 2018)	■	□	□	□	□	64.1	
(Zhang et al., 2019b)	■	□	□	□	□	66.6	
(Arnab, Doersch, and Zisserman, 2019)	□	■	□	□	■	77.8	54.3
(Jiang et al., 2019)	■	□	□	□	■	87.7	53.8
(Kanazawa et al., 2018)	■	□	□	□	■	88.0	58.1
(Iskakov et al., 2019)	□	□	■	□	□	17.7 ³	20.8
(Wang et al., 2019a)	■	□	□	□	□		40.7
(Shi et al., 2018)	■	□	□	□	□		43.7
(Kovalenko et al., 2019)	□	■	□	□	■		51.2
(Lee, Lee, and Lee, 2018)	■	□	□	□	□		52.8
(Zhao et al., 2019)	■	□	□	□	□		57.6
(Omran et al., 2018)	■	□	□	□	■		59.9
(Tian et al., 2019)	■	□	□	□	■		62.9
(Zhou et al., 2017)	■	□	□	□	□		64.9
(Chen et al., 2019a)	■	□	□	□	□		68.0
(Pavlakos et al., 2018)	■	□	□	□	■		75.9
(Pavlakos et al., 2019)	■	□	□	□	■		75.9

³Because of the annotation error in the Human3.6M dataset, a part of the evaluation data has been filtered out.

2.6 Conclusion

Convolutional neural networks have become the most dominant tool in human pose estimation because they were not dependent on the manually-designed representation and could find the representation besides the mapping to output. The state-of-the-art 3D human pose estimation methods trained and evaluated on the publicly available datasets (e.g., Human3.6M) were using multi-view images captured by several synchronized and calibrated cameras to estimate the three-dimensional position of joint centers.

These methods may potentially be utilized as the basis for developing an accurate, easy-to-use, and cost-effective marker-less motion capture system for clinical application. However, the current human pose dataset may not be applicable to train and assess the pose estimation methods for clinical gait study because of the limited number of subjects (there were only two subjects in the test set of the Human3.6M dataset), lack of pathological cases, and the errors introduced by marker placement on subjects' regular clothing. A new pose dataset, well adapted for gait study, should be collected.

The new dataset should consist of subjects of different age groups (e.g., children and young adults), pathological and asymptomatic subjects. The reference values for joint centers should be of high accuracy. A marker-based motion capture system may be used for obtaining the reference values. Indeed, the markers should be placed directly on the anatomical landmarks by professional individuals. Also, a medical imaging system may be used to improve the accuracy of the reference values of joint centers. The RGB cameras (the hardware of the marker-less motion capture system) should be calibrated and synchronized with the marker-based system. In the following chapter, we aim to collect the new pose dataset, well adapted for clinical gait study.

Chapter 3

ENSAM pose dataset

Herein, we aim to illustrate the collection of a new dataset, referred to as ENSAM pose dataset, well adapted for clinical gait study. A marker-less motion capture system will be proposed. Despite the Human3.6M dataset that four RGB cameras were placed at the four corners of the capture volume, the cameras' configuration will be changed so that the cameras' installation becomes simpler and more convenient.

In the Human3.6M dataset, the joint centers' reference positions were obtained by applying forward kinematics on the joint angles provided by the marker-based motion capture system's skeleton fitting procedure. To improve the accuracy of the reference joint centers, not only the markers will be placed on the anatomical landmarks by professional individuals but also a medical imaging system (EOS[®] system) will be used to reduce the errors attributed to marker placement. The bony segments (e.g., pelvis, femur, and tibia) and reflective markers will be reconstructed, and the anatomical landmarks (e.g., femur condyles) will be registered to the reflective markers.

In this chapter, we will explain all elements of the data collection and processing for the ENSAM pose dataset, experimental setups, calibration and synchronization procedures, annotation process, and the recorded walking trials. Several samples of the ENSAM pose dataset are shown in 3.1

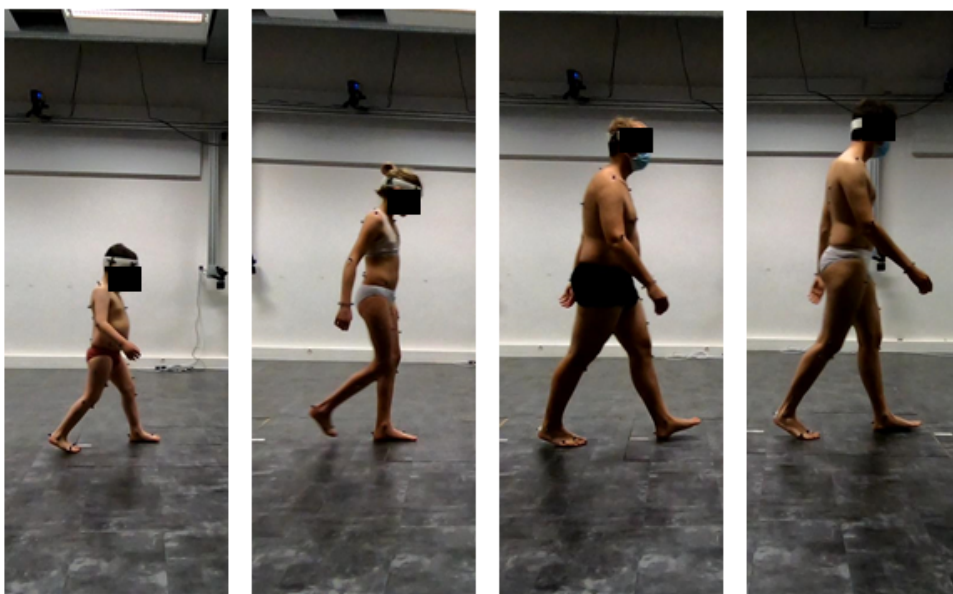


FIGURE 3.1: Samples from the ENSAM pose dataset.

3.1 Materials and methods

3.1.1 Reference marker-based motion capture system

The marker-based motion capture system is the Vicon **Motion Capture (MoCap)** system. The MoCap system's cameras are twelve cameras, including four Vicon Vero v2.2 cameras and eight Vicon Vero v1.3 cameras with a respective resolution of 2048x1088 pixels and 1280x1024 pixels. The frequency of data acquisition for data collection of the ENSAM pose dataset is 100 Hz. An image of the Vicon Vero camera is shown in Figure 3.2a.

The Vicon system is equipped with a device, which is called Lock sync box. This device is designed to integrate and synchronize third-party devices with the Vicon system. This device has 64 analog input channels, 8 GPO (General Purpose Output) programmable connections, and a few other connections. The analog input channels and GPO connections will be used to synchronize the marker-less MoCap system with the Vicon system.



FIGURE 3.2: (A) Vicon Vero v1.3 camera (B) Vicon calibration device placed on the floor to set the global coordinate system of the Vicon system.

3.1.2 Marker-less motion capture system

We designed the marker-less MoCap system with four RGB cameras. The camera is GoPro Hero 7 Black that can record videos with various video resolutions, frame rates, field of views, and aspect ratios. The selected specifications for the marker-less MoCap system is the resolution 1920x1080 pixels, frame rate 100 frames per second¹, linear field of view, and 16:9 aspect ratio. The selected frame rate was the same for the marker-based MoCap system.

We mounted each couple of cameras on a single aluminum bar (refer to Figure 3.3) and placed it on the subject's frontal and lateral view. This camera configuration was selected to make the installation and calibration easy and fast. For simplicity, as shown in Figure 3.4, the cameras installed on the **subject's frontal view** will be called **frontal cameras**, and the ones installed on the **subject's lateral view** will be called **lateral cameras**. As shown in Figure 3.3, a specific 3D printed fastening helps mount a single camera on the aluminum bar. On average, the distance and angle between every two cameras mounted on a single bar are 75 cm and 15°, respectively. Each aluminum bar is mounted on a tripod with a height of 1 m.

¹The frame rate is dependent on the video format. GoPro Hero 7 Black can record with two video formats, namely, NTSC and PAL. The video format is required when we aim to display the recorded videos on a TV. For example, the right format for Europe is PAL and for North America is NTSC. The appropriate selection of video format helps prevent flicker on a TV.



FIGURE 3.3: **Left:** the GoPro Hero 7 Black camera is mounted on the aluminum bar using the 3D printed fastening. The camera's position and orientation can be adjusted by the three degrees of freedom. **Right:** two cameras are mounted on the aluminum bar, which is installed on a tripod.

3.1.3 Experimental setup

Figure 3.4 shows the floor plan of the experimental setup. Twelve Vicon Vero cameras are mounted to the walls around the capture volume. The frontal cameras are placed on the subject's frontal view, the lateral cameras on the subject's lateral view. The global coordinate system is approximately in the middle of the capture space. The subject moves either toward the frontal cameras or the opposite direction. When the subject moves toward the frontal cameras, the subject's right side is seen by the lateral cameras. Moving in the opposite direction, and the lateral cameras see the subject's left side. Synchronization LEDs are placed on the ground so that the frontal and lateral cameras can see it. The longest path that the marker-less and marker-based MoCap systems can record the subject while moving is approximately 4 meters.

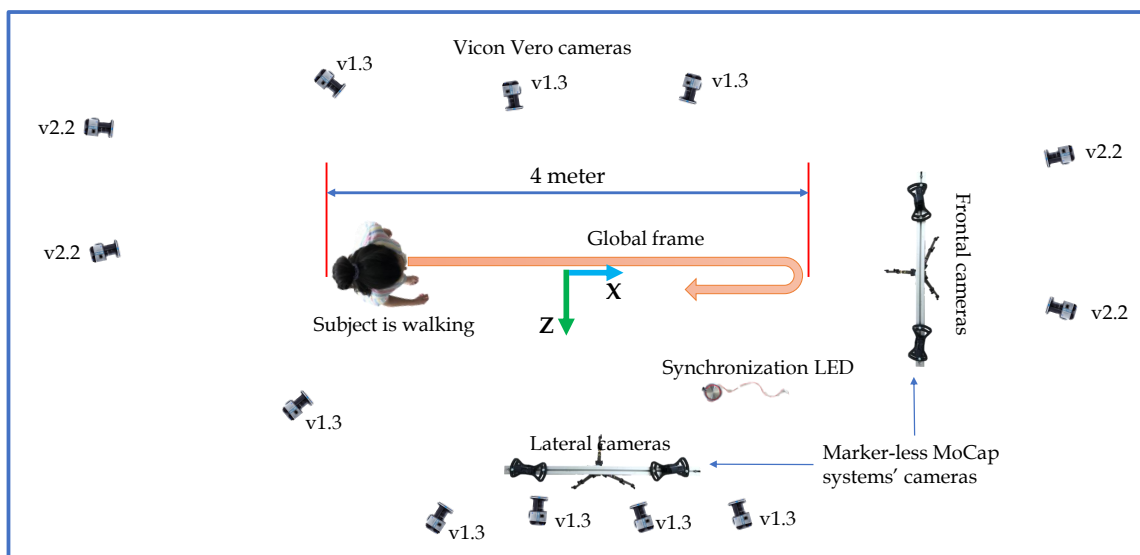


FIGURE 3.4: Floor plan showing placement of the marker-based MoCap systems' cameras (Vicon Vero) and the marker-less MoCap systems' cameras (GoPro).

3.1.4 Calibration

Reference marker-based motion capture system

A Vicon calibration device, as shown in Figure 3.2b, is used to calibrate the Vicon system. In the first step, the calibration device should be waved throughout the entire capture volume of the Vicon system, while being visible to the cameras. Each Vicon camera should record the calibration device for a minimum number of frames (e.g., 2000 frames). Then, the Vicon system calculates the calibration parameters for the Vicon cameras. In the second step, the calibration is adjusted on the floor, and the global coordinate system is set based on the position of the calibration device in the capture volume.

Marker-less motion capture system

The calibration device, shown in Figure 3.5a, consists of a checkerboard (9×14 , 40 mm) and six reflective markers. The checkerboard is purchased from calib.io, which is a commercial producer of calibration targets. Six markers are fastened to the checkerboard using 3D printed fasteners. Then, the marker's positions with respect to the checkerboard were measured. To this end, a biplanar radiography image, as shown in Figure 3.5b, was captured using the EOS[®] system. As shown in Figure 3.5c, the radiography helped 3D reconstruct the checkerboard and markers. Finally, the markers' 3D positions with respect to the checkerboard were computed.

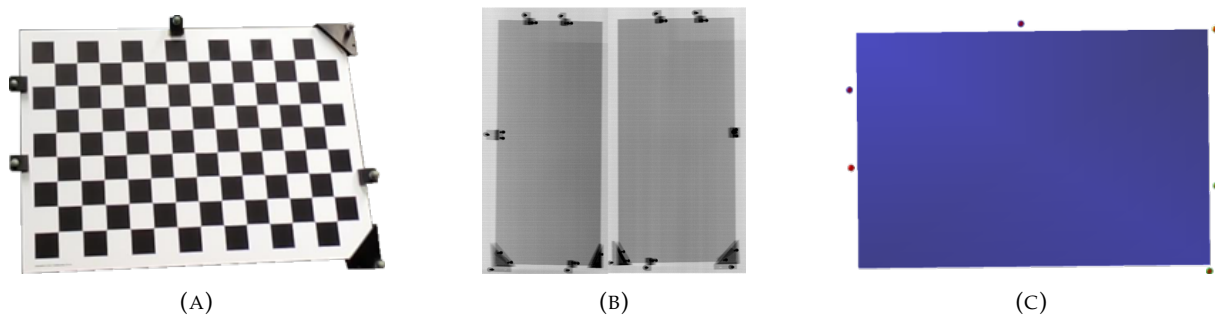


FIGURE 3.5: (A) The calibration device consisting of the checkerboard and six markers. (B) The biplanar radiography for measuring the position of the markers with respect to the checkerboard. (C) The 3D reconstructed checkerboard and markers using the biplanar radiography.

A developed procedure was applied to simultaneously calibrate the marker-less motion capture system and set the same global coordinate system for the marker-less and marker-based motion capture systems. The marker-less motion capture system consists of frontal and lateral cameras. To calibrate the marker-less motion capture system, we first apply the developed procedure to calibrate the frontal cameras, then the lateral cameras. The developed calibration procedure can be summarized in four main steps, including **data recording**, **data pre-processing**, **intrinsic and extrinsic calibration**, **setting the same coordinate system** for the marker-less and marker-based motion capture systems:

1. **Data recording:** The calibration device is placed at twenty different positions ($I = 20$). The marker-less motion capture system's frontal cameras record videos of the calibration device, and the marker-based motion capture system record the markers' positions. Figure 3.6 shows the placement of the calibration device at four different positions.
2. **Data pre-processing:** The recorded videos' frames are averaged to have one image for each position of the calibration device. The averaging also helps reduce the noises.

3. **Intrinsic and extrinsic calibration:** The camera model is a pinhole camera model (refer to Appendix A) with three radial and two tangential distortion parameters. The intrinsic (focal lengths, principal points, skew coefficient, and distortion parameters) and extrinsic (relative position and orientation between the two cameras) parameters are estimated using the MATLAB® Stereo Camera Calibrator App.
4. **Setting the same coordinate system:** The 3D positions of the checkerboard's corners are obtained using the triangulation method. And, the 3D positions of markers are computed thanks to the markers' known position on the checkerboard. Therefore, the 3D position of markers with respect to the frontal cameras' coordinate frame ($ML^i = \{ML_1^i, \dots, ML_6^i\}; i \in \{1, \dots, I = 20\}$) are computed. The 3D positions of the markers ($Mb^i = \{Mb_1^i, \dots, Mb_6^i\}; i \in \{1, \dots, I = 20\}$) were recorded by the marker-based MoCap system in the first step. Therefore, the markers' 3D positions are known in the marker-less and marker-based MoCap systems' coordinate systems. The rotation matrix R_f and the translation vector T_f between the frontal cameras of the marker-less MoCap system's and marker-based MoCap system's coordinate systems can be computed using Equation 3.1.

$$\min_{R_f, T_f} \sum_{i=1}^{I=20} \sum_{j=1}^{J=6} \|Mb_j^i - (R_f ML_j^i + T_f)\| \quad (3.1)$$

After the calibration of the frontal cameras, the procedure above is repeated for the lateral cameras to calibrate the entire marker-less MoCap system.



FIGURE 3.6: Calibration of the marker-less motion capture system: Placement of the calibration device at four different positions.

The accuracy of calibration is reported by two parameters, **reprojection** and **3D errors**. **Reprojection error**, which demonstrates the accuracy of intrinsic and extrinsic camera calibration, is the average distance between the detected calibration device keypoints (104 checkerboard corners) and the corresponding world keypoints projected into the image plane. For each position of the calibration device, the reprojection error across all checkerboard corners is computed. **3D error** is measured by $\|Mb_j^i - (R_f ML_j^i + T_f)\|; i \in \{1, \dots, 20\}, j \in \{1, \dots, 6\}$, which shows

the accuracy of setting the same coordinate system for marker-less and marker-based motion capture systems.

3.1.5 Synchronization

The GoPro Smart Remote controls the simultaneous recording of frontal and lateral cameras. However, the Smart Remote does not support the synchronized video recording, and the four GoPro cameras might start video recording with few seconds of delays with respect to each other. Also, the marker-less MoCap system should be synchronized with the marker-based MoCap system. Therefore, the objective is to synchronize all the marker-less MoCap system cameras together and the marker-less MoCap system with the marker-based one.

The marker-less MoCap system's cameras and the marker-based MoCap system record using the same frequency (100 Hz) of data acquisition. Therefore, the synchronization aims to find the temporal offset (the distance between two matching frames) between the recorded data. Once the temporal offset is obtained, and the recorded data are synchronized, each recorded frame by either marker-less or marker-based MoCap system refers to the same instance in the universal time.

The implemented synchronization technique is inspired by the study of (Shrestha et al., 2009), which synchronized a multi-camera setup using flash detection. We first explain the synchronization of the marker-less MoCap system's cameras and then the synchronization of the marker-less MoCap system with the marker-based one.

Synchronization of the marker-less MoCap system's cameras: A time-dependent signal turns ON and OFF a set of red LEDs placed in the marker-less MoCap system's capture space. Therefore, when the Marker-less MoCap system is recording videos, the LEDs are also being recorded in the videos, while becoming ON and OFF by the time-dependent signal. Then, the LEDs' status – being ON or OFF – in every frame of the recorded videos is determined using image processing techniques (**red color detection**). This procedure helps retrieve the time-dependent signal, which controlled the LEDs' status, for each marker-less MoCap system's camera. Finally, the cross-correlation between the retrieved signals from the marker-less MoCap system's cameras helps find the temporal offset between the recorded videos.

The **red color detection** is performed in three simple steps. First, a manually determined bounding box crop the LEDs in the image planes. Second, the image's color space is converted from RGB to HSV. The HSV color space uses three values – hue (color information), saturation (intensity or purity), and value (brightness) – to represent colors. Therefore, if the hue value of a pixel is within a specific known range, the color is detected as red. Third, if the number of red pixels in the cropped image is more than a specific threshold, the LED's status is detected as ON.

Synchronization of the marker-less MoCap system with the marker-based MoCap system: An Arduino UNO board, which is a microcontroller, generates the time-dependent signal to control the LEDs. The Arduino board is connected to the Vicon system using the Lock sync box. Whenever the Vicon system starts data recording, a constant signal (through the GPO connection) is sent to the Arduino board, and the Arduino board starts generating the time-dependent signal. The same time-dependent signal is also sent to the Lock sync box (through the analog input) and is recorded in the data. Then, once the Vicon system stops recording, the constant signal is cut off, and the Arduino board stops generating the time-dependent signal. Since the marker-based MoCap system also records the time-dependent signal, it helps synchronize the marker-less and marker-based MoCap systems.

3.1.6 Data collection

Subjects participated in this study after informed consent. The relevant ethics committee has approved the work (CPP 06036, CPP 06001, Paris VI). The subjects were asked to be with their underwear. Fifty-one reflective markers, as shown in Figure 3.7, were placed directly on the subjects' anatomical landmarks by orthopedic surgeons. For more information regarding the marker placement, refer to Appendix B.

Once the markers are placed on a subject's anatomical landmarks, a static acquisition is performed by the marker-based MoCap system. Figure 3.7 shows the static posture during the static acquisition. This static acquisition is essentially used for the post-processing of the data. A low-dose biplanar X-ray image is then taken with the EOS[®] system (EOS Imaging - France) in a standard standing posture (Chaïbi et al., 2012). This image is furtherly used to reconstruct the 3D bony segments' models that can be registered to the placed reflective markers on the body segments.

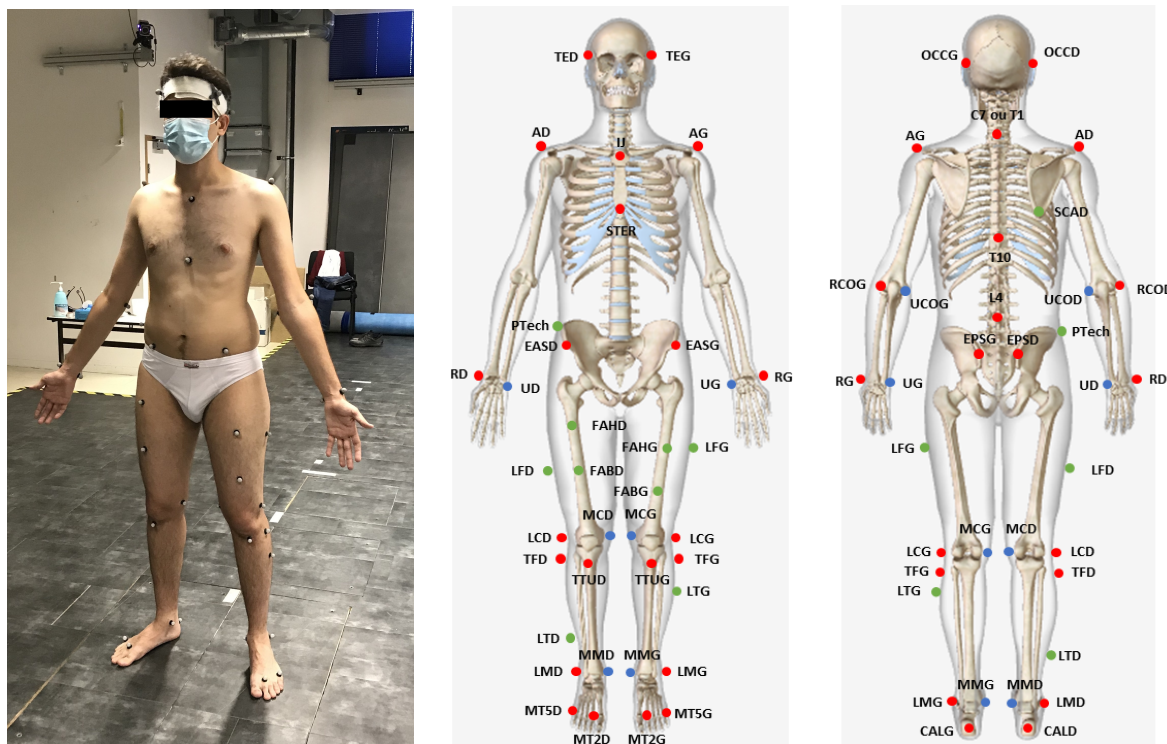


FIGURE 3.7: The image on the left: static posture; placement of fifty-one reflective markers on a subject's anatomical landmarks. The two images on the right: the marker-set used for the marker placement on the subjects' anatomical landmarks.

After the static and EOS acquisition, subjects were asked to walk for fifteen walking trials at their comfortable self-selected walking speed while being recorded by the marker-less and marker-based MoCap systems. Figure 3.8 shows a subject performing a walking trial. Also, the subjects performed 15 trials of the Timed-Up-and-Go-Test (TUG) (Podsiadlo and Richardson, 1991) consisting of standing up from a chair, walking forward for three meters, turn around, walking for three meters, and sitting down.



FIGURE 3.8: A subject performing a walking trial. This series of images are captured by a lateral camera of the marker-less MoCap system.

3.1.7 Annotation data

The annotation data consists of the **3D position of 16 joint or segment centers** (hereafter, for simplicity called joint centers) – head, neck, shoulders, elbows, wrists, trunk, pelvis, hips, knees, and ankles – and **subjects' bounding boxes in multi-view images** (four images recorded by the four cameras at the same time instant). The annotation data are obtained from the recorded data by the marker-based motion capture system and the EOS[®] system. First, we focus on extracting the reference values for the 3D position of joint centers, and then we determine the bounding boxes using the landmarks' positions.

Joint centers

The first step for processing the walking trials data, using the Vicon Nexus software, is to create a labeling skeleton template for the utilized marker-set. The labeling skeleton template consists of 13 segments (arms, forearms, head, trunk, pelvis, thighs, shanks, and feet) and 13 joints connecting the segments. Next, for each subject, using the created labeling skeleton template, the markers recorded during the static acquisitions, are reconstructed and manually labeled. This manual labeling helps create a subject-specific Vicon labeling skeleton that can be used for automatic labeling of the recorded walking trials.

After marker reconstruction, labeling, and gap-filling using the Vicon Nexus software, we process the bi-planar radiographs acquired by the EOS system. First, 3D shape models of the bony segments, including pelvis, femurs, tibias, and fibulas, are reconstructed (Mitton et al., 2006; Chaibi et al., 2012). Second, the reflective markers are reconstructed by fitting markers' sphere models to the markers contours visible on the bi-planar X-ray images. Figure 3.9 shows the 3D reconstruction of bony segments and reflective markers. These two steps help register the bony segments' anatomical frames, constructed using the bi-planar X-ray images, to the marker-based motion capture system's environment.

Herein, we define the centers for the 16 joints or segments. The upper limbs centers, including head, neck, shoulders, elbows, wrists, and trunk, are defined based on the markers in the marker-based MoCap environment. Table 3.1 shows the definition of the centers above. On the other hand, the centers of lower limbs, including pelvis, hips, knees, and ankles, are defined using the 3D reconstructed shape models. In other words, the lower limbs joint or segment centers are first defined in the 3D reconstructed body shapes, then registered in the marker-based MoCap system's environment.

The ankle joint center is defined as the barycenter of the tibias' malleolus. In Figure 3.9, the malleolus is shown in green color. For each knee joint center, two spheres, using the least square method, are fitted into the lateral and medial condyles of the femur's reconstructed mesh. Figure 3.9 shows the femur's condyles in green color. The centers of the fitted spheres represent the centers of the lateral and medial condyles. The knee joint center is defined as the mid-point between the lateral and medial condyles' centers. A similar approach is incorporated to define the hip joint center. A sphere is fitted to the femoral head, which is shown in Figure 3.9, using the least square method. The center of the fitted sphere representing the femoral head's center is defined as the hip joint center. Finally, the pelvis hip joint center is defined as the mid-point between the left and right center of two spheres fitted into the left and right acetabulum. Figure 3.9 shows the Pelvis' acetabulum in green color.

TABLE 3.1: The definitions of the upper-limbs' joints or segment centers

Joint or segment	Location	Markers
Head	center	TED, TEG, OCCD, OCCG
Neck	mid-point	IJ, C7
Left shoulder	coincident	AG
Right shoulder	coincident	AD
Left elbow	mid-point	UCOG, RCOG
Right elbow	mid-point	UCOD, RCOD
Left wrist	mid-point	RG, UG
Right wrist	mid-point	UD, RD
Trunk	coincident	T10

Bounding boxes

Once the position of the body joint centers is entirely determined in the marker-based motion capture environment, the joint centers are projected into the image planes of the marker-less motion capture system's cameras. Then, subjects' bounding boxes are determined using the projected joint centers. The bounding box is defined by four sides – top, bottom, left, and right. The top side of the bounding box is above the highest joint center, which usually is the head. The bottom side is below the lowest joint center, typically one of the ankles. The left and right sides are farther than the extreme left or extreme right joint centers.

If the bounding box is rectangular-shaped, it should be transformed into square-shaped, because the inputs to the convolutional neural networks are square-shaped. To this end, the bounding box's height and width are calculated, compared, and the greater value is selected as the side's length. Naturally, the rectangular-shaped bounding box's height is greater than its width. Then, the square-shaped bounding box is obtained using the bounding box's center and the selected side's length.

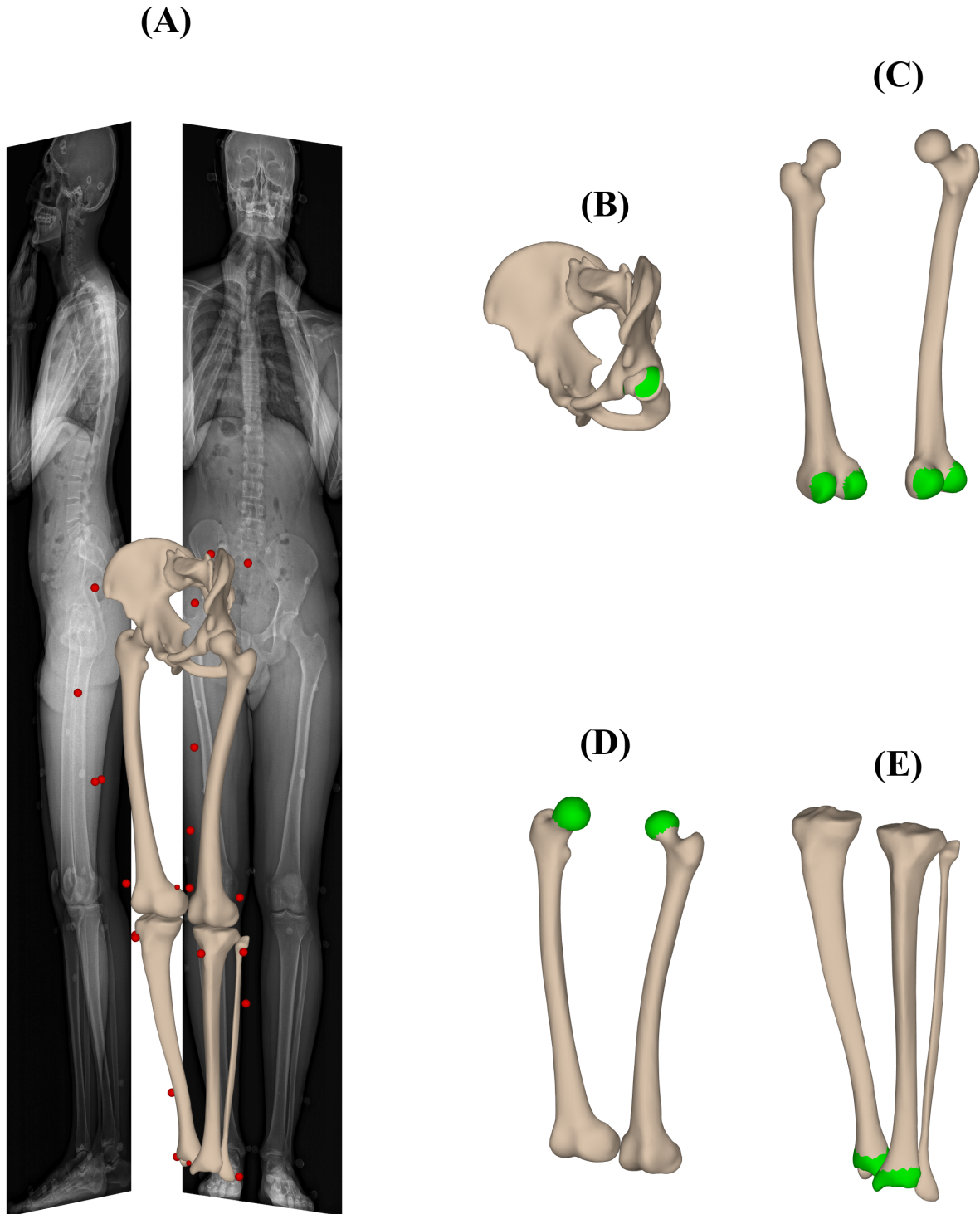


FIGURE 3.9: (A) The 3D reconstructions of pelvis, femurs, tibias, fibulas, and registered markers, using the biplanar X-ray image. (B) Pelvis 3D shape model; the green area indicates the acetabulum (cotyloid cavity). (C) Femurs 3D shape models; the green areas indicate the lateral and medial condyles. (D) Femurs 3D shape models; the green areas show the femoral heads. (E) Tibias 3D shape models; the green areas represent the malleolus.

3.1.8 Walking trials

Several conditions should be met to have a valid sample of the pose dataset. First, subjects should be in the common capture space of the marker-less and marker-based MoCap systems. Second, there should not be missing markers among the reconstructed markers of a walking trial. This phenomenon occurs when a subject is at the boundaries of the marker-based MoCap system's capture volume that some markers cannot be reconstructed because those markers could not be seen by a minimum number of cameras. Third, the bounding boxes should fit into the captured images by the frontal and lateral cameras. This phenomenon happens when a subject is at the boundaries of the marker-less MoCap system's capture volume that subject is inside the capture space, but the left or right side of the bounding box does not fit into the captured image. Therefore, the number of valid frames to form the pose dataset in every walking trials is less the captured frames by either marker-less or marker-based MoCap systems.

In section 3.1.6, we stated that every subject performed fifteen walking trials. Nevertheless, ten walking trials were selected to form the pose dataset. The selection was based on the quality of walking trials in terms of reconstruction, labeling, and gap-filling for markers in the recorded walking trials.

3.2 Results

3.2.1 Calibration

As explained in section 3.1.4, in each session of data collection, we calibrated the marker-based and marker-less MoCap systems. The marker-based system was calibrated following the proprietary guideline. Thus, the accuracy of its calibration is not discussed. Regarding the marker-less motion capture system, Figure 3.10 displays the placement of the calibration device at twenty different positions with respect to its frontal cameras. The colorful planes refer to the calibration device's checkerboard, and red spheres represent the six reflective markers.

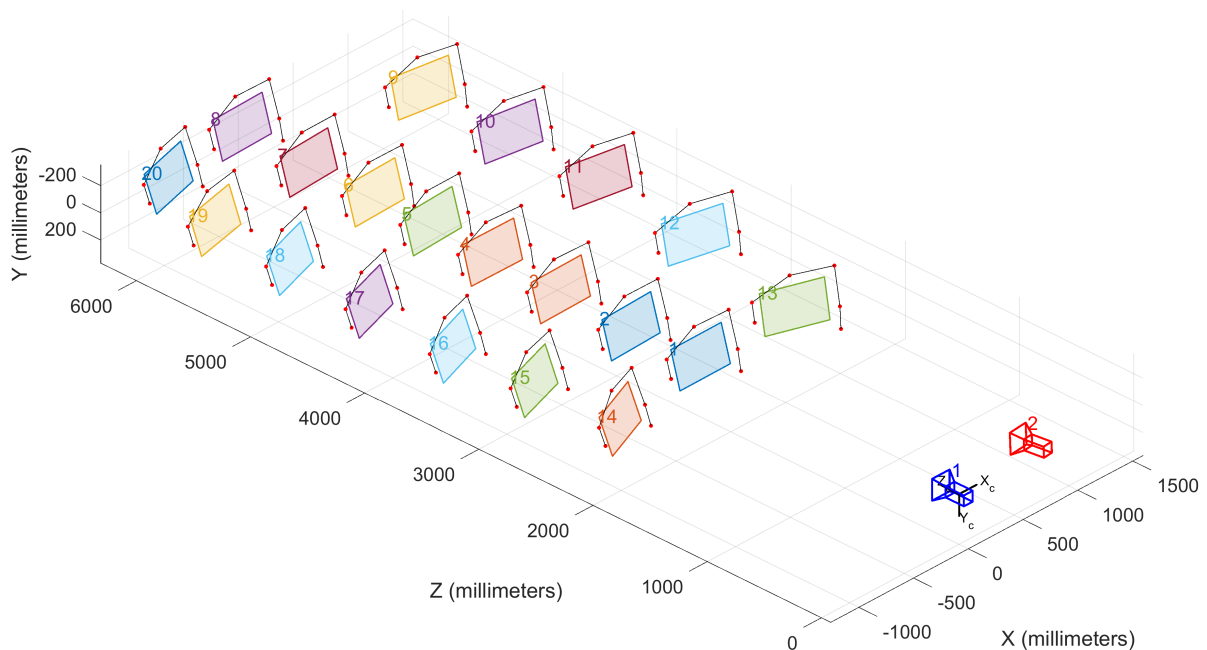


FIGURE 3.10: The calibration device placed at twenty different positions with respect to the frontal cameras in a typical session.

Figure 3.11 shows the reprojection error for every twenty positions of the calibration device. The overall mean reprojection error is 0.06 *pixels* (RMS: 0.07 *pixels*, range: 0.00 – 0.26 *pixels*).

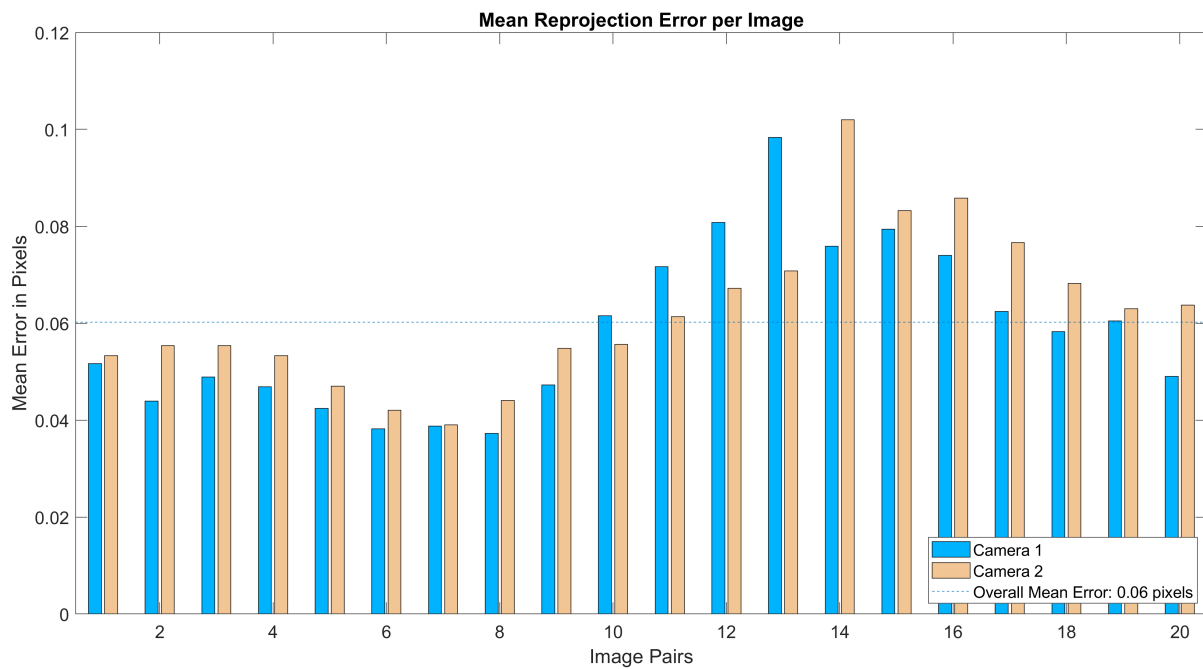


FIGURE 3.11: The mean projection error for captured images at different positions (as shown in Figure 3.10) with respect to the frontal cameras.

Figure 3.12 shows the results of setting the same global coordinate frame for the marker-less and marker-based MoCap systems. The 3D error represents the distance between the markers'

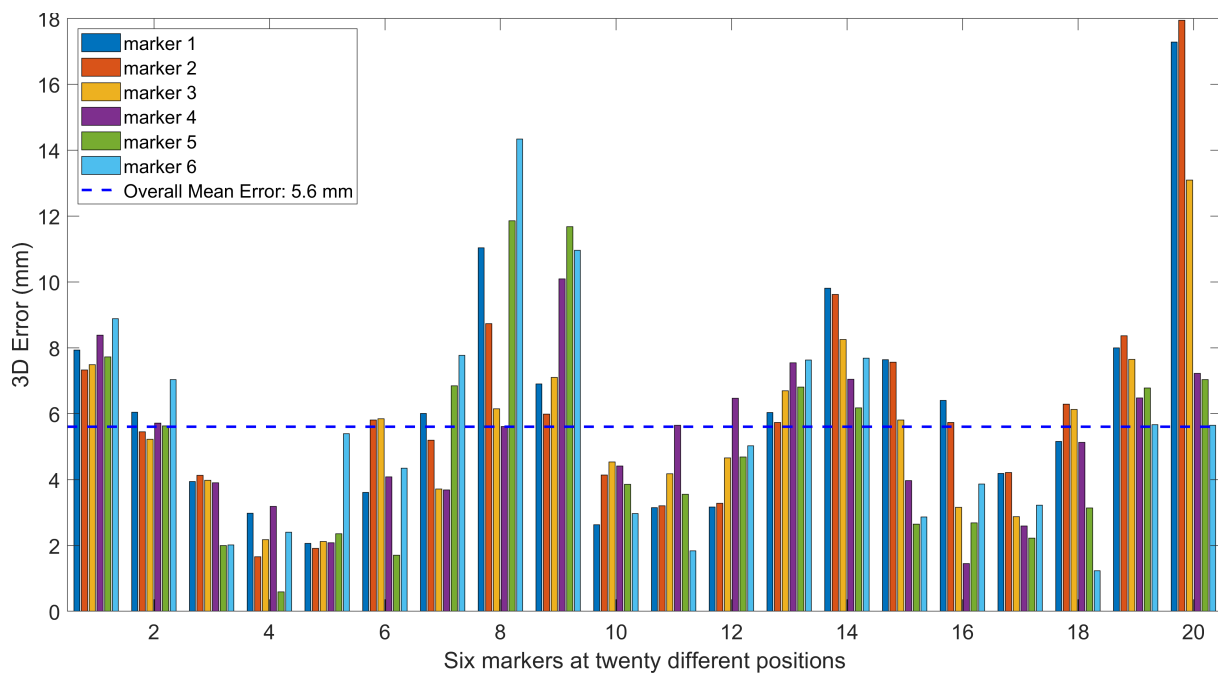


FIGURE 3.12: The 3D error for every marker at different positions (as shown in Figure 3.10) with respect to the frontal cameras.

position retrieved in the marker-based MoCap system's coordinate system and the markers' position retrieved in the marker-less MoCap system's frontal cameras' coordinate system transformed to the marker-based MoCap system's coordinate system. The overall mean 3D error is 5.6 mm (RMS: 6.4 mm , range: $0.6 - 17.9 \text{ mm}$). The mean error (standard deviation) along the axes of the global coordinate system is X-axis (anteroposterior): 0.00 mm (2.36), Y-axis (vertical): 0.00 mm (5.73), Z-axis (mediolateral): 0.00 mm (1.64), demonstrating zero bias.

Table 3.2 shows the reprojection error and 3D error of the marker-less MoCap system's frontal and lateral cameras' calibration across different data collection sessions. The mean 3D error, for frontal and lateral cameras, across all data collection sessions is 4.3 mm (RMS: 1.0 mm , range: $3.0 - 5.8 \text{ mm}$). It should be noted that once the marker-less system is calibrated, and its cameras have not been moved, the marker-less system can record across different gait analysis sessions.

TABLE 3.2: The calibration error of the marker-less motion capture system across the sessions of data collection

session	Frontal cameras				Lateral cameras			
	reproj. err. ^(a) (pixels)		3D err. ^(b) (mm)		reproj. err. ^(a) (pixels)		3D err. ^(b) (mm)	
	Mean	RMS	Mean	RMS	Mean	RMS	Mean	RMS
01	0.07	0.08	3.4	3.8	0.08	0.09	2.6	2.9
02	0.08	0.11	3.3	4.6	0.07	0.08	3.8	4.2
03	0.08	0.10	4.6	5.4	0.08	0.10	4.1	4.6
04	0.11	0.13	6.5	7.2	0.08	0.09	4.2	4.8
05	0.06	0.07	4.7	5.4	0.07	0.08	3.8	4.2
06	0.06	0.08	4.3	4.9	0.07	0.08	6.0	6.5
07	0.07	0.08	7.5	8.2	0.07	0.08	3.2	3.6
08	0.06	0.07	5.0	5.5	0.07	0.08	3.7	5.6
09	0.06	0.07	4.1	4.4	0.07	0.08	3.6	3.9
10	0.06	0.07	4.3	4.8	0.07	0.08	3.8	4.2
11	0.06	0.07	3.9	4.3	0.07	0.08	3.7	4.1
12	0.06	0.08	7.8	8.8	0.06	0.07	3.4	3.6
13	0.06	0.07	6.5	7.4	0.07	0.08	3.1	3.5
14	0.06	0.08	3.6	4.2	0.07	0.08	4.3	4.7
15	0.05	0.06	3.9	4.3	0.07	0.08	3.2	3.6
16	0.06	0.07	2.9	3.1	0.07	0.08	3.1	3.4
17	0.06	0.07	3.5	3.9	0.08	0.10	3.7	4.2
18	0.07	0.08	6.6	7.4	0.09	0.10	4.9	5.5
19	0.06	0.07	5.0	5.5	0.07	0.08	3.6	4.1
20	0.06	0.07	5.6	6.2	0.08	0.09	3.4	3.8
21	0.06	0.07	5.9	6.5	0.08	0.09	3.6	4.0
22	0.06	0.07	4.7	5.2	0.08	0.09	3.3	3.8
23	0.06	0.07	3.4	3.8	0.08	0.09	3.8	4.2

^(a) reprojection error

^(b) 3D error

3.2.2 Synchronization

Figure 3.13 shows the time-dependent signal recorded by frontal cameras in a walking trial across 800 frames. There is a temporal offset of 6 frames (60 milliseconds) between the recorded signals. This calculated temporal offset helps synchronize the recorded data by two frontal cameras.

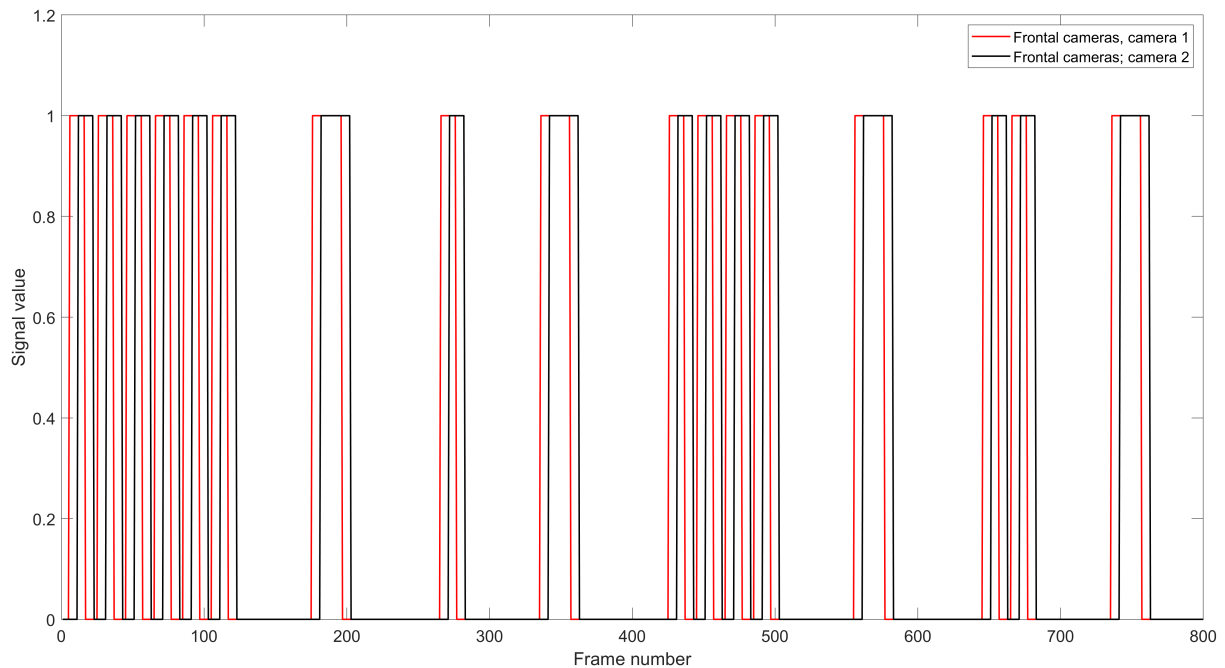


FIGURE 3.13: The time-dependent synchronization signal recorded by the frontal cameras of the marker-less motion capture system.

3.2.3 Population

Thirty-one subjects (19 males and 12 females) participated in this study. Subjects were on average 25 years old (SD: 9, range: 6–44), mean height was 168 cm (SD: 18 cm, range: 125 – 199 cm), body mass was 65 kg (SD: 17 kg, range: 30 – 90 kg), and Body Mass Index (BMI) was 22.5 kg/m² (SD: 3.0 kg/m², range: 16.9 – 29.4 kg/m²). Table 3.3 shows the subjects' demographic data, including age, height, weight, BMI, gender and pathology if applicable.

TABLE 3.3: The demographic data of subjects who participated in the collection of the ENSAM pose dataset.

subj.	age (y.)	height (cm)	body mass (kg)	BMI (kg/m ²)	gender	pathology
01	13	152	62	26.8	F	XLH ^(a)
02	13	130	30	17.8	F	XLH
03	11	125	35	22.4	F	XLH
04	15	165	51	18.7	F	XLH
05	12	148	37	16.9	F	XLH
06	12	137	32	17.0	F	XLH
07	6	133	31	17.5	M	XLH
08	31	162	63	24.0	M	-

Continued on next page

Table 3.3 – Continued from previous page

subj.	age (y.)	height (cm)	body mass (kg)	BMI (kg/m ²)	gender	pathology
09	26	172	65	22.0	M	-
10	25	173	72	24.1	M	-
11	25	175	65	21.2	M	-
12	27	182	75	22.6	M	-
13	22	172	65	22.0	M	-
14	23	186	70	20.2	M	-
15	26	184	75	22.2	M	spondylolisthesis ^(b)
16	25	185	86	25.1	M	-
17	24	188	80	22.6	M	-
18	27	181	71	21.7	M	-
19	25	160	60	23.4	F	-
20	25	175	90	29.4	M	-
21	29	187	73	20.9	M	-
22	25	173	64	21.4	F	-
23	31	171	76	26.0	M	-
24	41	159	55	21.8	F	moderate scoliosis ^(c)
25	31	199	90	22.7	M	-
26	29	167	58	20.8	F	-
27	32	165	68	25.0	F	-
28	34	175	81	26.4	M	- ^(d)
29	44	172	73	24.7	M	-
30	39	160	65	25.4	F	-
31	29	186	85	24.6	M	-

^(a) X-Linked Hypophosphatemia is an inherited disorder that causes rachitic deformities of the lower limbs.

^(b) Spondylolisthesis is a spine disorder in which a vertebra (spine bone) moves forward onto the bone below it.

^(c) Scoliosis is the 3D deformation of the spine.

^(d) An operation had been performed on the knee anterior cruciate ligament of this subject.

3.2.4 Data processing

Figure 3.14 shows the number of frames of each subject's walking trials. The total number of frames across all subjects' walking trials is 93331.

Figure 3.15 shows a sample of the ENSAM pose dataset – a multi-view image captured by the frontal and lateral cameras of the marker-less motion capture system, sixteen body joint centers projected on the image planes, and four square-shaped bounding boxes, obtained from the marker-based MoCap system.

3.3 Discussion

In this chapter, we illustrated the designed marker-less motion capture system, the calibration, and synchronization procedures. Then, we collected the ENSAM pose dataset to evaluate the designed marker-less motion capture against a marker-based motion capture system in terms of clinically relevant gait parameters.

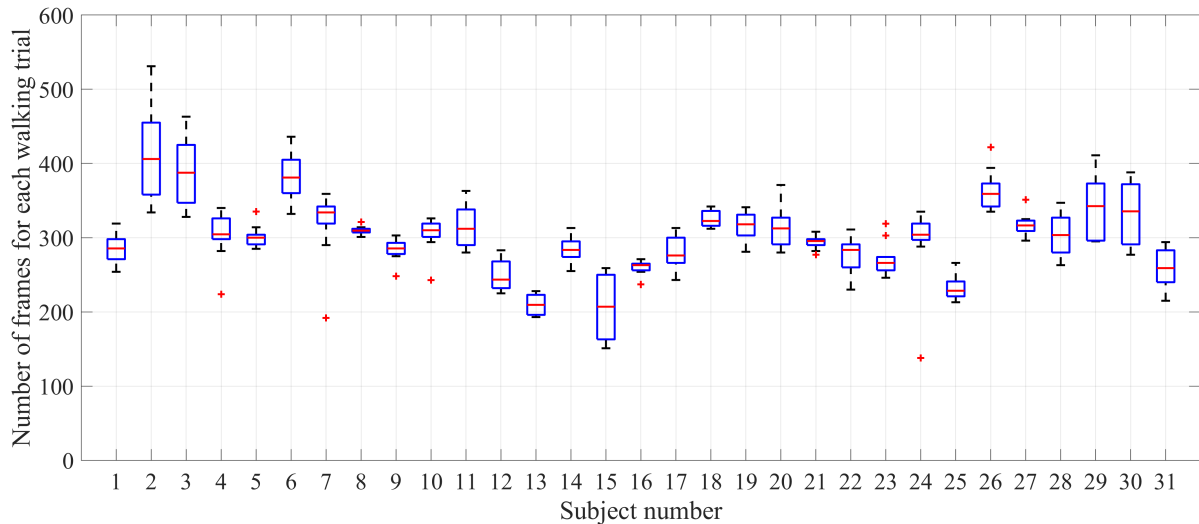


FIGURE 3.14: Number of frames of subjects' walking trials

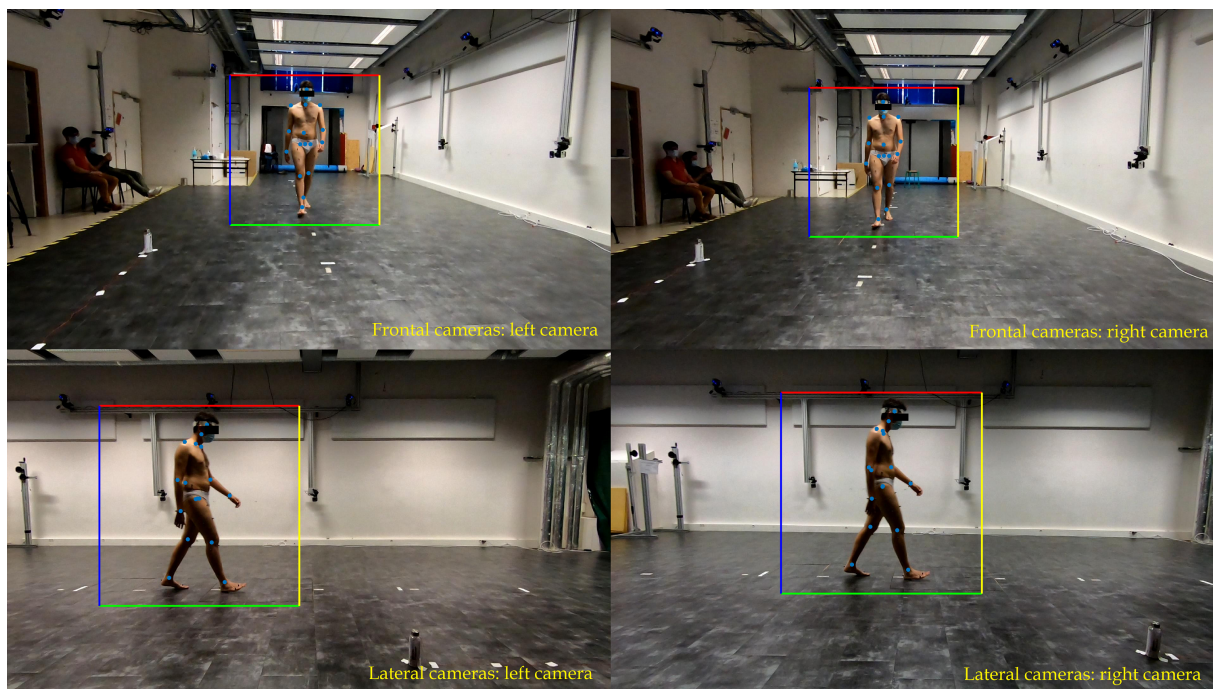


FIGURE 3.15: The projected joint centers on the four image planes of the marker-less MoCap system's cameras; The subject's square-shaped bounding boxes retrieved for every image plane.

3.3.1 Experimental setup

In chapter 1, we studied commercial marker-less MoCap systems; and most of the commercial MoCap systems based on RGB cameras used eight cameras. Reducing the number of cameras helps lessen the installation time and costs of materials. (Iskakov et al., 2019) helped us choose four number of cameras to design the marker-less motion capture system. (Iskakov et al., 2019) evaluated the proposed human pose estimation method's performance on the CMU Panoptic dataset (Joo et al., 2017), collected by more than thirty HD cameras, versus a different number of cameras. The results showed that increasing the number of cameras (from two to twenty-eight) would decrease errors. Nevertheless, a considerable improvement was observed merely

until four cameras, and after four cameras, the rate of improvement decreased significantly. Therefore, the designed marker-less MoCap system consists of four cameras.

The four-camera setup could be arranged in various ways. For example, consider a rectangular capture space, four cameras can be installed on the four corners or sides. Nevertheless, two factors contributed mainly to the design of the four-camera setup. First, the installation of four cameras should be fast and straightforward. Second, the cameras should be installed on different views to gather more information. Therefore, instead of mounting every camera on a single tripod and placing it in a unique position, every two cameras (frontal and lateral cameras) were installed on a tripod using an aluminum bar. The approximate distance of 75 *cm* between either frontal or lateral cameras was considered to gather more information. The approximate angle of 15° between the two cameras was also considered to increase the marker-less MoCap system's capture volume, especially for the cameras installed on the subject's lateral view. Further studies would be necessary to find the optimal camera configuration for the marker-less motion capture system.

The GoPro Hero 7 cameras can record in different video modes in terms of resolutions, the field of views, and frame rates. Concerning the field of view, there are three fields of view available when recording video, namely, linear, wide, and super view. The linear field of view removes the fisheye effect. Indeed, the fisheye cameras can record with a wide-angle view. However, compared to the pinhole camera model, which is suitable for the linear field of view, the camera calibration model becomes more complex and less accurate for stereo vision (Tu et al., 2013; Ohashi et al., 2016). Also, selecting the same frame rate as the marker-based MoCap system (100 Hz) would make the synchronization more straightforward. Hence, if we consider the constraints on the field of view and frame rate, the only possible video resolution was 1920×1080 *pixels*, adequate for further MoCap analysis.

One of the design criteria for the marker-less motion capture system was being low-cost and, more precisely, less than 5,000 €. The price of each GoPro Hero 7 Black camera is about 350 €, and the price of the Smart Remote is nearby 100 €. Therefore, the total cost of the cameras and the remote is around 1,500 €. Later, in Chapter 4, we will discuss the total costs of the designed marker-less MoCap system.

3.3.2 Calibration

The calibration procedure aims to calibrate the marker-less motion capture system and set the same global coordinate system as of the marker-based motion capture system. The intrinsic and extrinsic parameters of either frontal or lateral cameras of the marker-less motion capture system are estimated using the MATLAB Stereo Camera Calibrator App. The mean reprojection error varied from 0.06 – 0.10 *pixels*. However, the reprojection error may not be intuitive to demonstrate the accuracy of the calibration. In a preliminary test, we first calibrated the frontal cameras using a printed checkerboard (14×20 , 20 *mm*). Then, we measured the checkers' sides width of another checkerboard (5×7 , 70 *mm*), placed at 35 different positions covering the entire capture space, by the calibrated frontal cameras. The RMS error of measuring 17 side width, in each place of the checkerboard, varied from 0.2 *mm* to 2.2 *mm* (mean RMS error = 0.8 *mm*). The results showed that the intrinsic and extrinsic calibration of the frontal cameras is precise enough to pursue the objective of this study.

The mean 3D error, which represents the accuracy of setting the same global coordinate system for the marker-based and marker-less MoCap systems, varied from 3.0 – 5.8 *mm*. Even

though this error is not restrictive for comparing the marker-less MoCap system against the marker-based system for measuring clinically relevant gait parameters, the error could be regarded as a limitation of this study.

The error of setting the same coordinate system can be attributed to various sources. To identify the sources of error, we first review the procedure of setting the same coordinate system for the marker-less and marker-based MoCap systems. The position of the calibration device's markers with respect to the marker-less motion capture system's frontal or lateral cameras' coordinate frames are retrieved in three steps. First, using a corner extraction algorithm, the 2D coordinates of the checkerboard's corners are obtained in image coordinates. Second, using the geometric triangulation, the 3D position of checkerboard corners is obtained. Third, thanks to the known geometric relations between the checkerboard and the markers of the calibration device, the position of the calibration device's markers are retrieved. The position of the calibration device's markers is recorded directly by the marker-based MoCap system. Hence, the errors can be attributed to the corner extraction algorithm, the triangulation, the 3D reconstruction of the calibration device which helped retrieve the position of markers with respect to the checkerboard, and the accuracy of the marker-based MoCap system.

There is a statistically significant difference (Two-Sample t-test) between the calibration error of the frontal and lateral cameras ($p < 0.05$). On average, the calibration error of the frontal cameras (mean: 4.8 mm, SD: 1.4 mm) is higher than the lateral cameras (mean: 3.7 mm, SD: 0.7 mm) because the depth of the capture volume is higher (approximately 7 m vs. 5 m). When the calibration device is placed further, the resolution of the calibration device in the captured image reduces, and subsequently, the accuracy of the corner detection algorithm decreases. This causes the reduced calibration accuracy of frontal cameras compared with the lateral cameras.

3.3.3 Synchronization

Marker-less motion capture system (GoPro cameras) can start and stop video recording by the capture button, voice command, or remote control. The only practical way to control the four cameras is by using a remote control. Nevertheless, the remote control does not support the synchronous recording of the GoPro cameras. Therefore, we used a set of LEDs controlled by a time-dependent signal to synchronize the marker-less MoCap system's cameras.

To ensure the synchronization quality, for several frames in every recorded walking trial, as shown in Figure 3.16, the position of markers captured by the marker-based MoCap system was projected into the image planes captured by the marker-less MoCap system. Then, the consistency between the image and projections of markers was checked.

3.3.4 Data processing

The number of frames of walking trials depends on several factors. The first factor is the annotation data. Sometimes, some markers' positions cannot be retrieved by the marker-based MoCap system because the markers cannot be seen by a minimum number of cameras. The second factor is the walking speed. Since the subjects were asked to walk at their comfortable walking speed, some subjects walked faster than the other subjects. When a subject walks faster, the number of recorded frames reduces. The third factor is the subjects' height. When a subject is taller, the bounding box is also larger. Thus, the subject reaches faster to the image's boundaries, and the length of the walking trial (number of frames) reduces.

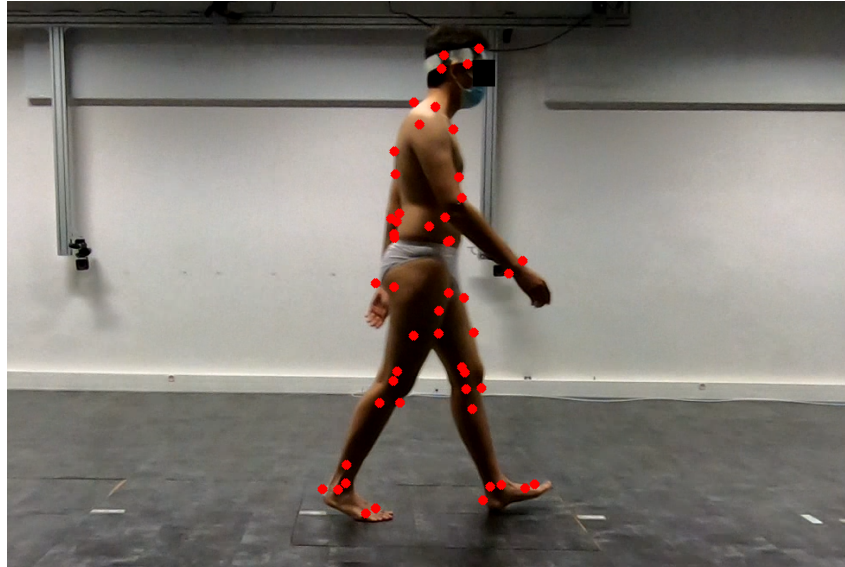


FIGURE 3.16: Projection of markers (captured by the marker-based motion capture system) into the image plane of a lateral camera of the marker-less motion capture system to ensure synchronization quality.

3.4 Conclusion

ENSAM pose dataset is designed and collected to assess the accuracy of the designed marker-less MoCap system against the marker-based MoCap systems for clinical gait study. The marker-less MoCap system consists of two pairs of cameras – called frontal cameras and lateral cameras – installed on two sides of the capture space. The marker-based MoCap system is a Vicon system equipped with twelve Vicon Vero cameras installed on the four sides of the capture space. The procedures for calibration and synchronization of these systems are explained comprehensively in this chapter.

Thirty-one volunteers participated in the data collection of this study after informed consent. Fifty-one markers were placed on the subject's body segments by orthopedic surgeons. The EOS system took a biplanar X-ray image. The biplanar X-ray image helped reconstruct the lower limb bony segments, determine joint or segment centers' accurate position, and register the determined centers to the reconstructed markers. Then, every subject performed fifteen walking trials while being recorded by synchronized and calibrated marker-less and marker-based MoCap system. Ten walking trials were further processed to be included in the ENSAM pose dataset.

The annotation data consists of joint centers and subjects' bounding boxes in image frames. The position of sixteen joint centers is retrieved using the marker-based and EOS systems for every subject's walking trials. The joint centers were projected to the image planes of the marker-less MoCap system's cameras using the calibration parameters. The square-shaped bounding boxes of subjects were obtained using the projected joint centers.

This chapter explained all the ENSAM pose dataset elements, including objectives, experimental setups, data collection, and data processing. In the following chapter, we use the ENSAM pose dataset to assess the designed marker-less MoCap system's accuracy against the marker-based MoCap system.

Chapter 4

Marker-less motion capture system: validity and reliability

The designed marker-less motion capture system should be accurate, cost-effective, and easy-to-use for **clinical gait analysis**. Therefore, we aim to evaluate the system in terms of clinically relevant gait parameters. In this chapter, we will evaluate the designed system on the ENSAM pose dataset. The dataset will be split into two independent sets – training and test sets. The marker-less motion capture’s human pose estimation method will be trained on the training set. The test set will then be utilized to assess the accuracy of the marker-less motion capture system against a marker-based motion capture system in terms of **joint position error**, **spatiotemporal gait parameters**, and **kinematic gait parameters**.

The designed marker-less motion capture system records a subject’s movement using four synchronized and calibrated RGB cameras. Then, the human pose estimation method (Iskakov et al., 2019), integrated into the marker-less motion capture system, estimates the position of sixteen joint centers (head, neck, shoulders, elbows, wrists, trunk, pelvis, hips, knees, and ankles) for every multi-view image of the captured sequence.

We first explain the human pose estimation methods and the training strategy. Then, the gait events detection method, which determines the time instants of heel strikes and toe-off events in walking trials, is illustrated. The gait events provide the basis for the calculation of the spatiotemporal and kinematic gait parameters. Finally, the spatiotemporal and kinematic gait parameters are computed and compared with the marker-based motion capture system.

4.1 Materials and methods

4.1.1 Human pose estimation

The implemented human pose estimation method was presented by (Iskakov et al., 2019). Iskakov et al. presented two approaches for human pose estimation: algebraic triangulation and volumetric triangulation. Algebraic triangulation was used mainly to estimate the pelvis center; then, volumetric triangulation was used for the other joint centers.

The outlines of the human pose estimation approach based on algebraic triangulation is shown in Figure 4.1. The input is a set of RGB images captured by a set of synchronized and calibrated cameras. The 2D backbone helps retrieve the 2D joint heatmaps and joints’ confidences. The ResNet-152 convolutional neural network (introduced in section 2.5.1), followed by a series of transposed convolutions, and a 1×1 -kernel convolutional neural network estimate the 2D joint heatmaps. Also, the ResNet-152 network, followed by a convolutional network, which consists of two convolutional layers, global average pooling, and three fully connected layers, estimate the joints’ confidences. The joints’ confidences define the reliability

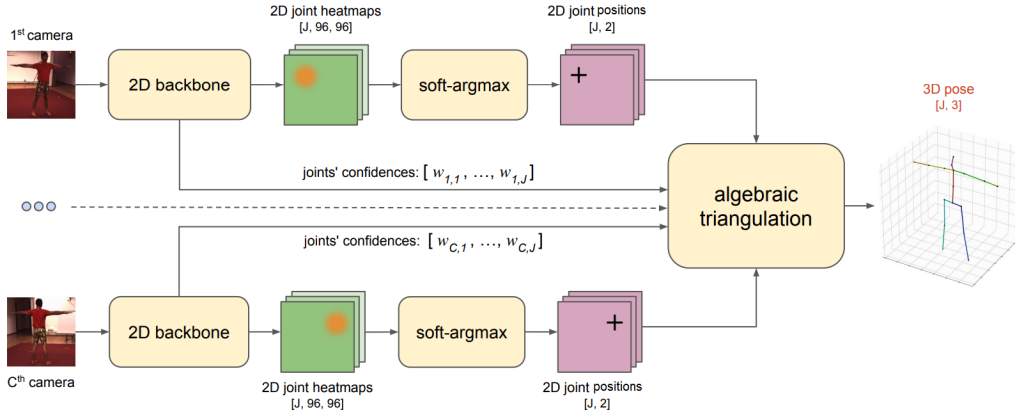


FIGURE 4.1: The human pose estimation method based on algebraic triangulation (Iskakov et al., 2019). J and C represent the joint and camera.

of the 2D joint heatmaps. The soft-argmax calculates the 2D joint positions based on the 2D joint heatmaps. First, the softmax of the 2D joint heatmaps across the spatial axes are computed. The softmax function normalizes the 2D heatmaps joints so that the heatmaps can be interpreted as probabilities. Then, the center of mass of the normalized 2D joint heatmaps (also referred to as soft-argmax) is calculated. The input to the algebraic triangulation is the 2D joint positions and joints' confidences, and the output is 3D joint positions.

The outlines of the human pose estimation method based on volumetric triangulation is shown in Figure 4.2. The 2D backbone consists of the ResNet-152 convolutional neural network, followed by a series of transposed convolutions and a single layer convolutional neural network with 1×1 kernel and $K = 32$ outputs. The 2D backbone output is called feature maps instead of joint heatmaps because the number of output channels does not represent the number of joints. The estimated feature maps are unprojected into 3D volumes. A $2.5m \times 2.5m \times 2.5m$ -sized volume, discretized into $64 \times 64 \times 64$ voxels, is placed around the pelvis center, which is estimated by the algebraic triangulation. The feature maps from different views are unprojected into the voxels using the camera calibration parameters. Then, the volumetric maps from all views are aggregated. The aggregated volumetric maps are then processed by a convolutional neural network, which has a similar structure to V2V-PoseNet

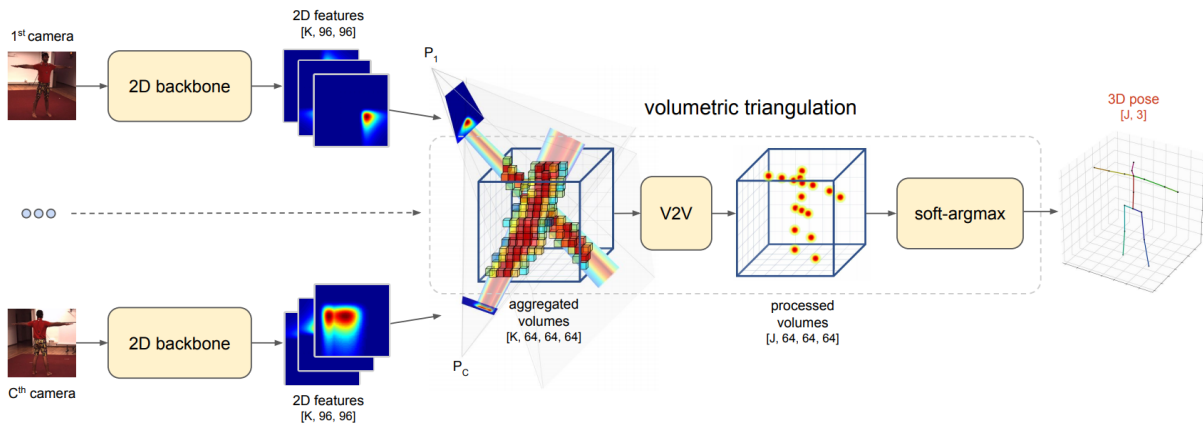


FIGURE 4.2: The human pose estimation method based on volumetric triangulation (Iskakov et al., 2019). C , K , and J represent the camera, number of output channels of the 2D backbone, and joint.

(Moon, Yong Chang, and Mu Lee, 2018), to produce 3D joint heatmaps. Similar to the algebraic triangulation, the 3D joint heatmaps are normalized by the softmax function. The 3D joint positions are then obtained by calculating the center of mass of the normalized 3D joint heatmaps.

The loss function of the algebraic triangulation was a robust per-joint Mean Squared Error (MSE). The loss function of the volumetric triangulation consists of a per-joint Mean Absolute Error (MAE) and a second term, which aimed to maximize the prediction for the voxel containing the reference value of the joint center.

4.1.2 Training and test of the human pose estimation method

The algebraic and volumetric triangulation methods were trained end-to-end on the Human3.6M dataset (Ionescu et al., 2014) (refer to section 2.1). The 2D backbone had already been pre-trained on the COCO dataset (Lin et al., 2014) and then finetuned on the MPII (Andriluka et al., 2014) (refer to section 2.1) and Human3.6M dataset.

We finetuned the algebraic and volumetric triangulation methods on the training set of the ENSAM pose dataset. The ENSAM pose dataset was randomly split into the train and test sets. The training set consisted of 57880 multi-view images – 19 subjects (S02, S03, S05, S06, S09, S10, S11, S12, S13, S14, S16, S17, S18, S19, S22, S24, S27, S30, and S31). And the test set consisted of 35451 multi-view images – 12 subjects (S01, S04, S07, S08, S15, S20, S21, S23, S25, S26, S28, and S29). The demographic data of the training and test sets is shown in Table 4.1.

TABLE 4.1: Demographic data of the training and test sets of the ENSAM dataset

	Training set	Test set
Subjects	19 subjects (9 Female, 10 Male)	12 subjects (3 Female, 9 Male)
Age (y.)	Mean: 24 (SD: 8, range: 11 – 41)	Mean: 26 (SD: 10, range: 6 – 44)
Height (cm)	Mean: 166 (SD: 19, range: 125 – 188)	Mean: 170 (SD: 17, range: 133 – 199)
Weight (kg)	Mean: 62 (SD: 17, range: 30 – 86)	Mean: 68 (SD: 17, range: 31 – 90)
BMI (kg/m²)	Mean: 22.0 (SD: 2.5, range: 16.9 – 25.4)	Mean: 23.3 (SD: 3.5, range: 17.5 – 29.3)
Pathology	4 XLH, 1 Scoliosis	3 XLH, 1 Spondylolisthesis

No data augmentation was applied during the finetuning of the algebraic or volumetric triangulation. The algebraic triangulation was finetuned for 22 epochs using the Adam optimizer with 10^{-4} learning rate. The volumetric triangulation was finetuned for 7 epochs using the same optimizer with 10^{-4} learning rate for the 2D backbone and a different learning rate 10^{-3} learning rate for the volumetric backbone.

First, we estimated the joint centers for the ENSAM test set, using the human pose estimation method, which was already trained on the Human3.6M dataset. Then, once the algebraic and volumetric triangulation methods were finetuned on the training set of the ENSAM pose dataset, first, the algebraic triangulation, then the volumetric triangulation, estimated the position of joint centers for the test set of the ENSAM pose dataset. The joint position errors were computed as the Euclidean distance between the estimated and reference joint positions. Also, the estimated joint positions were further used to estimate the spatiotemporal and kinematic gait parameters.

4.1.3 Detection of gait events

The gait event detection algorithm was adapted from (Zeni Jr, Richards, and Higginson, 2008) and (O'Connor et al., 2007). (Zeni Jr, Richards, and Higginson, 2008) proposed a velocity-based algorithm to detect the gait events for the marker-based motion capture systems. In this algorithm, first, the sacral marker's coordinate is subtracted from the heel marker's coordinate at every frame of walking trials. Second, the heel maker's velocity is computed by taking the first derivative of the coordinates using finite difference equations. Thus, the heel strike is the instant at which the X component of hip velocity sign changes from positive to negative; The toe-off is the instant at which the X component of hip velocity sign changes from negative to positive. This algorithm was modified to become adapted for the marker-less motion capture system.

The modified algorithm was utilized to detect gait events for marker-less and marker-based motion capture systems. In this algorithm, the ankles and pelvis joint centers were first low pass filtered using a zero-phase fourth-order Butterworth filter with the cutoff frequency 7 Hz (O'Connor et al., 2007). Second, the pelvis' coordinate was subtracted from the ankles' coordinate at every walking trial frame. Third, the ankles' velocity was computed by taking the first derivative of the coordinates using finite difference equations. Fourth, the heel strike and toe-off events were detected based on the ankle's relative velocity. At heel strike, the X component of the ankle's relative velocity changed from positive to negative; At toe-off, the X component of velocity changed from negative to positive.

4.1.4 Spatiotemporal gait parameters

The measured spatiotemporal gait parameters are gait speed, stride length, step length, step width, step time, stance time, swing time, and cadence. The definitions of all spatiotemporal gait parameters are given in Table 4.2.

TABLE 4.2: The definitions of the spatiotemporal gait parameters

Parameter	Definition
Stride length (<i>cm</i>)	The distance between the ankle joint position at the heel strike and the same ankle joint position at the next heel strike along the direction of progression.
Gait speed (<i>m/s</i>)	The measured stride length divided by the measured stride time.
Step length (<i>cm</i>)	The distance between the ankle joint position at the heel strike and the opposite ankle joint position at the next heel strike along the direction of progression.
Step width (<i>cm</i>)	The distance, perpendicular to the direction of progression, between the ankle joint position at the heel strike and the opposite ankle joint position at the subsequent heel strike.
Step time (<i>sec</i>)	The elapsed time between the heel strike of one foot and the subsequent heel strike of the opposite foot.
Stance time (<i>sec</i>)	The elapsed time between the heel strike of one foot and the subsequent toe-off of the same foot.
Swing time (<i>sec</i>)	The elapsed time between the toe-off of one foot and the subsequent heel strike of the same foot.
Cadence (<i>steps/sec</i>)	Number of steps per second; measured by one second divided by the measured step time.

The basic spatiotemporal gait parameters, including stride length, step length, step width, and the association of these parameters with gait events (heel strike and toe-off), are illustrated in Figure 4.3.

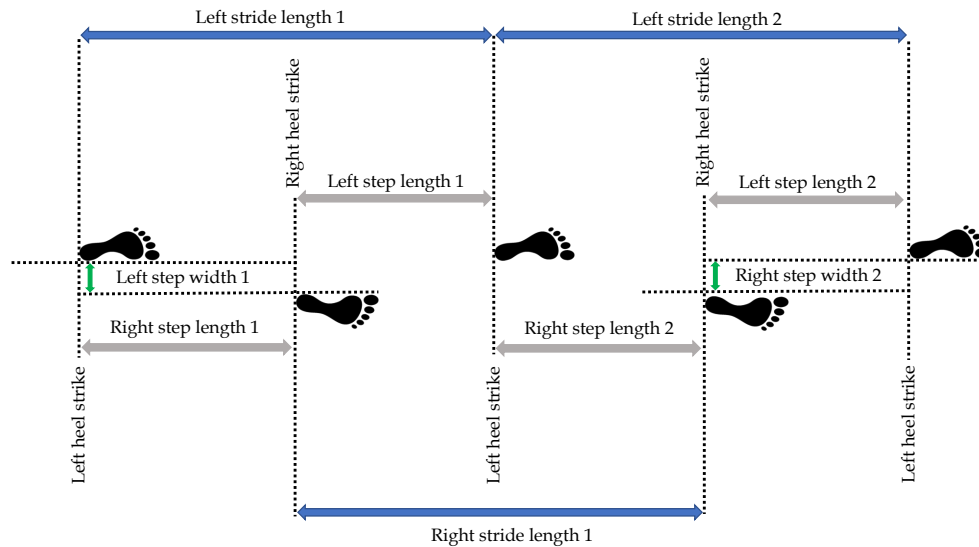


FIGURE 4.3: The definition of the stride length, step length, and step width.

4.1.5 Kinematic gait parameters

The measured kinematic gait parameters can be classified into joint center-based and coordinate system-based. First, the joint center-based kinematic gait parameters (trunk flexion, femur abduction-adduction, femur extension-flexion, and joint center-based knee extension-flexion) are introduced in Table 4.3.

The coordinate system-based kinematic gait parameters are hip extension-flexion, hip abduction-adduction, knee extension-flexion, pelvis rotation, and pelvis abduction-adduction. The hip joint kinematics describes the movement between the pelvis and femur bony segments; the knee joint kinematics describes the movement between the femur and tibia bony segments. The pelvis kinematics demonstrates the pelvis movement with respect to the global coordinate system. The coordinate systems of the pelvis, femurs, and tibias, based on the reconstructed bony segments by the bi-planar X-ray images, were defined according to Table 4.4. The kinematic gait parameters (pelvis, hip, and knee joints kinematics) were calculated following the procedure explained in Section 1.1.1.

The marker-less motion capture system estimated two points for each body segment – femur: hip and knee joints, tibia: knee and ankle joints, pelvis: hip joint centers. Thus, for each body segment, only one axis of the coordinate system could be formed. The other axis was retrieved using a priori knowledge to form a coordinate system. To this end, the gait analysis data of sixty-six asymptomatic subjects, independent from the ENSAM pose dataset, were used to compute the average rotations of bony segments about the axes of their coordinate systems in a gait cycle. The gait analysis data were collected by a marker-based motion capture system and EOS system. For instance, in a gait cycle, for the pelvis coordinate system, the Z-axis was

TABLE 4.3: The definitions of the joint center-based kinematic gait parameters

Paremeter	Definition
Trunk flexion	The angle between two vectors. The first vector is formed by joining the pelvis center to the neck joint center, and the second vector is the global coordinate system's vertical axis.
Femur abduction-adduction	The angle measured in the laboratory frontal plane between two vectors. The first vector is formed by joining the knee and hip joint centers, projected into the laboratory frontal plane. The second vector is the global coordinate system's vertical axis.
Femur extension-flexion	The angle measured in the laboratory sagittal plane between two vectors. The first vector is formed by joining the knee and hip joint centers, projected into the laboratory sagittal plane. The second vector is the global coordinate system's vertical axis.
JB* knee extension-flexion	The angle between two vectors. The first vector is formed by joining the knee and hip joint centers. The second vector is formed by joining the knee and ankle joint centers.

JB*: Joint center-Based

the line connecting the left and right hip joint centers, pointing to the right. The Y-axis or X-axis could be retrieved using a priori knowledge to form a coordinate system.

4.1.6 Reliability of kinematic gait parameters

The inter-trial reliability of kinematic gait parameters was assessed using the method proposed by (Schwartz, Trost, and Werve, 2004). First, the mean kinematic gait parameters across all

TABLE 4.4: Definitions of the pelvis, femur, and tibia coordinate systems

Segment	Axis	Definition
Pelvis	Z	The line connecting the centers of pelvis' left and right acetabulum, pointing to the right.
	X	The line perpendicular to the plane defined by the Z-axis and the line connecting the center of the pelvis' sacral plate and the origin of the pelvis, pointing anteriorly.
	Y	The line perpendicular to the X-axis and Z-axis, pointing cranially.
Femur	Y	The line connecting the origin of the femur and the center of the femoral head, pointing cranially.
	Z	The line connecting the center of the femur's medial and lateral condyles projected into the plane perpendicular to the Y-axis.
	X	The line perpendicular to the Y-axis and Z-axis, pointing anteriorly.
Tibia	Y	The line connecting the barycenter of the distal tibia and the intersection of the diaphyseal tibial axis and the tibial plateau, pointing cranially.
	Z	The line connecting the center of the tibia's medial and lateral condyles projected into the plane perpendicular to the Y-axis
	X	The line perpendicular to the X-axis and Z-axis, pointing anteriorly.

the walking trials was calculated. Then, the differences of the kinematic gait parameters at all walking trials with respect to the mean was computed. Thus, the inter-trial reliability was obtained by computing the standard deviation of the differences computed for all the walking trials. Consider $\Phi_m^{subj}(t)$ denotes a kinematic gait parameter (e.g., knee extension-flexion), for a subject (*subj*), performing a walking trial (*m*) at a time instant (*t*). $\Delta\Phi_m^{subj, trial}(t)$ represents the inter-trial deviations:

$$\Delta\Phi_m^{subj, trial}(t) = \Phi_m^{subj}(t) - \frac{1}{N_{trials}} \sum_{m=1}^{N_{trials}} \Phi_m^{subj}(t) \quad (4.1)$$

After the concatenation of the deviations across different subjects, the estimated standard error can be computed using Equation 4.2 (Schwartz, Trost, and Wervey, 2004).

$$\sigma_{\Phi(t)}^{source} = \sqrt{\frac{1}{N_{total} - 1} \sum_{p=1}^{N_{total}} (\Delta\Phi^{source})^2} \quad (4.2)$$

4.2 Results

First, the human pose estimation was trained on the Human3.6M dataset and evaluated on the ENSAM test set. Then, the human pose estimation method was finetuned on the ENSAM training set and evaluated on the ENSAM test set. The difference between the estimated joint centers and the corresponding reference values, along the axes of the global coordinate system, is presented in Table 4.5. The mean differences (biases) were relatively high (within 115.3 mm) and became closer to zero (within 7.4 mm) across all joints after finetuning on the ENSAM training set. Also, the 95% confidence intervals (2 standard deviations) were significantly reduced across all joints. For instance, along the Z-axis for the hip joint, the confidence interval was reduced from 106.3 mm to 19.8 mm.

TABLE 4.5: Joint position error (*mm*) across all subjects' walking trials along X, Y, and Z axes of the global coordinate system.

Axis		ankles	knees	hips	pelvis	trunk	neck	head	wrists	elb. ^(e)	sho. ^(s)
Trained on the Human3.6M dataset											
X	mean	3.9	3.9	4.3	6.0	0.5	6.5	0.7	3.8	5.5	5.1
	2SD	68.4	50.4	38.0	23.8	155.0	25.4	42.0	54.8	67.3	42.6
Y	mean	25.6	-30.7	-49.4	-50.2	-115.3	-20.4	-14.8	-5.0	-19.4	35.0
	2SD	36.8	47.1	36.6	29.2	132.1	43.9	40.7	37.9	33.0	20.7
Z	mean	4.0	3.1	-0.4	0.0	-1.3	1.4	-3.8	5.1	2.7	-0.5
	2SD	30.3	29.3	106.3	11.6	23.9	12.0	14.0	98.7	36.6	75.0
Trained on the Human3.6M dataset and finetuned on the ENSAM training set											
X	mean	0.0	-0.1	0.1	0.3	-1.1	-0.2	0.1	-0.2	0.5	0.0
	2SD	13.2	16.4	26.3	18.1	29.9	18.4	31.4	27.0	34.5	31.8
Y	mean	-0.2	-1.2	-1.9	-3.5	-7.4	1.8	-0.2	1.6	-1.4	0.0
	2SD	6.5	14.1	22.0	22.1	48.1	14.9	2.9	20.9	13.8	15.6
Z	mean	-0.2	0.0	0.4	0.3	0.5	0.3	0.1	-1.1	0.5	0.2
	2SD	5.1	8.4	19.8	7.8	13.3	8.7	9.7	24.8	16.9	50.4

^(e)elbows, ^(s)shoulders

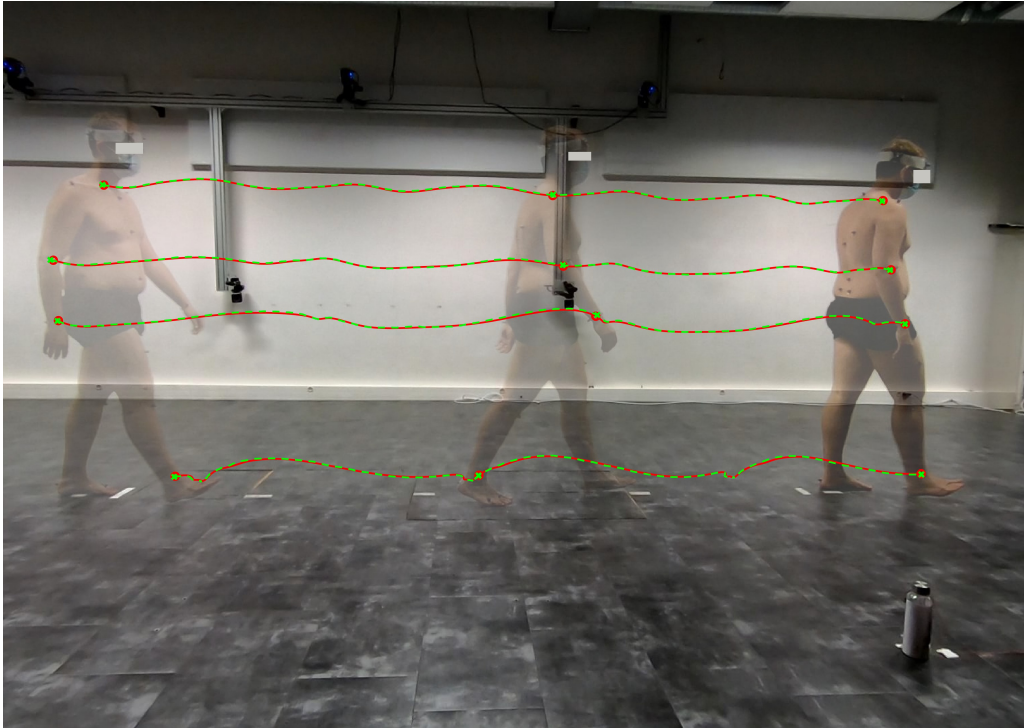


FIGURE 4.4: The joint center's projection on the image plane of the lateral camera of the marker-less motion capture system. Trajectories, from top to bottom: neck, right elbow, right wrist, and right ankle. Red trajectories: by the marker-based system. Green trajectories: by the marker-less system.

After finetuning on the ENSAM training set, the ankles, elbows, shoulders, and head position errors along axes were close to zero (the absolute value less than 0.5 mm), demonstrating zero bias. The knees, hips, wrists, and neck position error along axes were less than 2 mm . However, the pelvis and trunk position error along the Y-axis (laboratory's vertical axis, as shown in Figure 3.4) were -3.5 mm and -7.4 mm , respectively. **The following results are reported only for after finetuning on the ENSAM training set.**

The comparison between the evaluation of the algebraic and volumetric triangulation methods on the ENSAM pose dataset's test set showed no statistically significant difference in joints position error ($p = 0.91$). However, since the volumetric triangulation was slightly more accurate than the algebraic triangulation (MPJPE, $15.3\text{ mm} < 15.5\text{ mm}$), the results are reported only for the volumetric triangulation method.

Figure 4.4 shows a subject performing a walking trial. The trajectories of the neck, right wrist, right elbow, and right ankle are projected on the image plane of the lateral camera of the marker-less motion capture system. The red trajectory is obtained by the marker-based motion capture system, whereas the green trajectory by the marker-less motion capture system. As this figure illustrates, the difference between the trajectories is minimal.

The joint position errors are shown in Table 4.6. There was no statistically significant difference among the mean joint position error of subjects ($p=0.95$). Also, there was no statistically significant difference between the mean joint position error of pathological subjects (S01, S04, S07, S15) and asymptomatic subjects (S08, S20, S21, S23, S25, S26, S28, S29) ($p=0.94$). Of the sixteen joint centers, as illustrated in Figure 4.5), the ankles were the most accurate (mean error = 6.1 mm). The second and third most accurate joints were the knees (mean error = 10.3 mm)

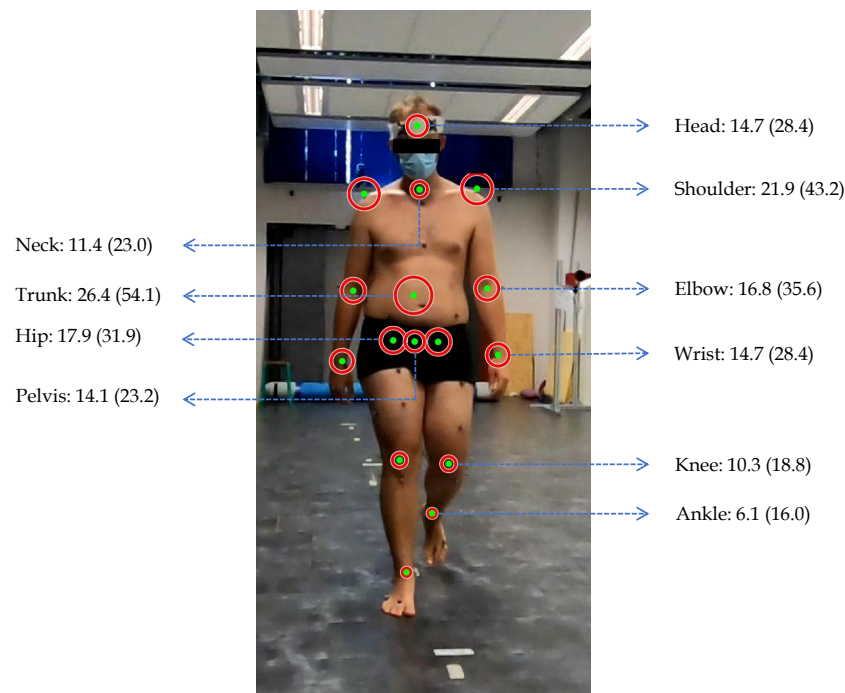


FIGURE 4.5: Graphical illustration of the mean joint position error across all subjects' walking trials. The diameter of circles represents a scaled magnitude of the mean joint position errors. The reported values are **Mean (95th percentile) mm**.

and neck (mean error = 11.4 mm), respectively. The least accurate joint center was the trunk (mean error = 26.4 mm).

The 95th percentile for the shoulder position error of subject 04 was 240.4 mm, which was high compared to other subjects' shoulder position error. This high value was due to several frames in three out of ten walking trials in which the shoulder positions were not estimated correctly. These three walking trials were excluded for the computation of gait kinematic parameters.

The kinematic gait events were determined independently from the marker-less motion capture system's estimated joint centers and the marker-based motion capture system's reference joint centers. The offsets between the gait events obtained by the marker-less and marker-based motion capture systems are shown in Table 4.7. A total number of 299 right heel strikes, 317 left heel strikes, 313 right toe-off, and 301 left toe-off events were determined from the 120 walking trials. The maximum offset was two frames across all gait events. Since marker-less and marker-based motion capture systems' frame rate was 100 Hz, the maximum offset corresponds to 20 milliseconds. 99% of gait events differences were within 1 frame (10 milliseconds). The Average frame offset for the gait events was close to zero (less than 1 millisecond).

The difference between the spatiotemporal parameters measured by the marker-less and marker-based motion capture systems are shown in Table 4.8. The biases between the spatiotemporal gait parameters measured by the marker-less and marker-based motion capture systems were close to zero (smaller than 0.1). The maximum absolute difference in gait speed is 0.014 m/sec. The step length and stride length have a similar upper limit of agreement (1.506 and 1.437 cm) and lower limit of agreement (-1.625 and -1.613 cm). The limits of agreements for

TABLE 4.6: Joint position error (*mm*) across subject's all walking trials

subject		ankles	knees	hips	pelvis	trunk	neck	head	wrists	elb. ^(e)	sho. ^(s)
01	mean	7.6	9.8	12.4	6.5	20.0	10.1	15.0	36.8	11.8	20.7
	prc. ^(r)	19.2	16.1	20.6	11.6	30.3	15.3	20.8	83.6	25.5	30.4
04	mean	5.2	11.4	23.5	17.3	12.5	11.1	21.4	16.9	12.3	32.5
	prc.	14.0	19.0	41.0	22.2	17.1	15.0	30.5	31.7	21.0	240.4
07	mean	7.0	12.9	14.0	8.5	52.7	7.8	10.6	12.4	10.0	23.7
	prc.	17.7	22.7	23.6	14.5	61.6	14.0	21.8	22.6	18.7	34.3
08	mean	5.4	11.9	24.1	15.7	12.3	8.1	13.8	19.7	22.5	26.2
	prc.	13.0	19.3	39.5	21.8	19.0	13.7	17.7	41.3	34.5	39.8
15	mean	5.4	10.5	20.7	14.8	48.0	7.2	11.4	10.3	9.2	16.6
	prc.	13.0	17.1	29.2	22.6	57.1	11.3	17.3	27.7	17.4	23.5
20	mean	7.0	11.6	17.8	13.8	22.2	6.4	16.6	11.8	8.6	12.8
	prc.	18.8	19.9	29.0	19.9	31.5	12.1	26.0	26.6	20.5	22.7
21	mean	5.2	7.5	20.6	18.6	15.1	9.7	18.6	10.4	15.7	18.2
	prc.	11.6	12.8	28.0	24.3	23.9	11.9	25.9	22.7	36.9	33.2
23	mean	7.0	7.6	13.7	10.8	14.4	16.6	6.6	10.2	22.4	15.0
	prc.	17.4	14.6	21.9	17.7	21.7	24.9	12.2	24.2	36.6	24.5
25	mean	6.9	9.2	24.9	19.6	27.6	22.2	15.1	17.4	20.9	23.4
	prc.	17.1	16.5	34.2	25.9	40.3	28.5	38.0	30.5	38.2	35.5
26	mean	6.1	9.8	14.0	13.1	27.7	7.8	26.6	14.8	11.5	36.7
	prc.	17.0	15.9	21.1	20.0	34.5	12.8	36.7	27.8	19.1	55.6
28	mean	6.2	12.6	22.2	19.8	23.0	13.9	10.6	15.4	25.6	16.5
	prc.	16.9	20.7	33.7	28.3	28.6	22.7	17.9	23.2	39.5	29.6
29	mean	5.1	8.9	10.7	12.4	42.7	17.4	7.0	15.2	30.1	15.9
	prc.	13.7	23.7	19.2	18.6	54.4	24.6	16.6	20.2	44.4	22.2
All	mean	6.1	10.3	17.9	14.1	26.4	11.4	14.7	16.0	16.8	21.9
	prc.	16.0	18.8	31.9	23.2	54.1	23.0	28.4	36.0	35.6	43.2

(e) elbows, (s) shoulders, (r) 95th percentile.TABLE 4.7: The difference (*frames*) between the kinematic gait events detected based on the marker-based and marker-less motion capture systems

Parameter	MD ¹	SD ²	LLoA ³	ULoA ⁴	MxAD ⁵	RMSD ⁶
right heel strike	0.07	0.60	-1.10	1.24	2	0.60
left heel strike	0.08	0.56	-1.03	1.18	1	0.57
right toe off	-0.03	0.67	-1.34	1.28	2	0.67
left toe off	0.00	0.77	-1.52	1.51	2	0.77

¹Mean Difference, ²Standard Deviation, ³Bland-Altman Lower Limit of Agreement, ⁴Bland-Altman Upper limit of agreement, ⁵Maximum Absolute Difference, ⁶Root Mean Square Difference.

the step width are approximately (-0.713 and 0.720 *cm*) half either the step length's or the stride length's limits of agreements. The maximum absolute difference of step, stance, and swing time is 20 *milliseconds*. The limits of agreement of cadence are -0.04 and 0.04 *steps/second*.

The difference between the kinematic gait parameters across 366 strides obtained by the marker-less and marker-based motion capture systems are shown in Table 4.9. The mean difference for trunk flexion (0.1°), femur extension-flexion (-0.2°), femur abduction-adduction (-0.2°), and joint center-based knee extension-flexion (0.0°) was close to zero. The mean difference for

TABLE 4.8: The difference between the spatiotemporal gait parameters obtained based on the marker-based and marker-less motion capture systems.

Parameter	MD ¹	SD ²	LLoA ³	ULoA ⁴	MxAD ⁵	RMSD ⁶	P-value ⁷
Gait speed (m/s)	0.001	0.003	-0.007	0.006	0.014	0.003	0.94
Stride length (cm)	-0.088	0.778	-1.613	1.437	2.071	0.782	0.93
Step length (cm)	-0.060	0.799	-1.625	1.506	2.651	0.800	0.90
Step width (cm)	0.004	0.366	-0.713	0.720	1.355	0.365	0.99
Step time (sec)	0.000	0.006	-0.013	0.012	0.020	0.006	0.94
Stance time (sec)	-0.001	0.008	-0.016	0.014	0.020	0.008	0.79
Swing time (sec)	0.001	0.007	-0.014	0.015	0.020	0.007	0.73
Cadence (steps/sec)	0.001	0.021	-0.041	0.042	0.091	0.021	0.94

¹Mean Difference, ²Standard Deviation, ³Bland-Altman Lower Limit of Agreement, ⁴Bland-Altman Upper limit of agreement, ⁵Maximum Absolute Difference, ⁶Root Mean Square Difference. ⁷Two-samples t-test p-values.

pelvis abduction-adduction (-0.5°) and hip ad-adduction (0.7°) was not close to zero but less than one degree. However, the mean difference for hip extension-flexion (2.6°), pelvis rotation (3.0°), and knee extension-flexion (-3.4°) demonstrated a bias which was more than 2.5° . The range of motion was averaged over all subjects' walking trials. Considering the normalized RMS difference (the root mean square difference divided by the range of motion), joint center-based knee extension-flexion was the most accurate kinematic gait parameter; The knee and femur extension-flexion parameters were the second most accurate parameters. The least accurate kinematic gait parameters were pelvis abduction-adduction and rotation angles.

TABLE 4.9: The difference (degrees) between the kinematic gait parameters obtained based on the marker-based and marker-less motion capture systems.

	Parameter	RoM ¹	RMS ²	MAD ³	MD ⁴	SD ⁵	LLoA ⁶	ULoA ⁷
Trunk	Flexion	5.8	1.4	1.1	0.1	1.4	-2.6	2.8
Pelvis	Ab-Adduction	3.8	2.8	2.1	-0.5	2.8	-5.9	4.9
	Rotation	7.1	4.1	3.4	3.0	2.8	-2.4	8.4
Hip	Extension-Flexion	31.6	8.4	6.6	2.6	7.9	-13.0	18.2
	Ab-Adduction	9.2	3.5	2.5	0.7	3.4	-5.9	7.4
Femur	Extension-Flexion	23.5	1.9	1.5	-0.2	1.9	-3.9	3.6
	Ab-Adduction	5.8	1.3	1.0	-0.2	1.3	-2.8	2.4
Knee	Extension-Flexion	57.1	4.5	3.8	-3.4	3.0	-9.3	2.5
	JB ⁸ Extension-Flexion	56.2	2.5	1.9	0.0	2.5	-4.8	4.8

¹Range of Motion, ²Root Mean Square Difference, ³Maximum Absolute Difference, ⁴Mean Difference, ⁵Standard Deviation, ⁶Bland-Altman Lower Limit of Agreement, ⁷Bland-Altman Upper limit of agreement, ⁸Joint center-Based.

The inter-trial error in gait kinematic parameters, obtained by either marker-less or marker-based motion capture systems, were computed for every point in the gait cycle. The reliability of a gait kinematic parameter was measured by the mean standard deviation of the inter-trial errors. The reliability of kinematic gait parameters obtained by the marker-less and marker-based motion capture systems is shown in Table 4.10. The reliability of gait kinematic parameters was almost the same in marker-less and marker-based motion capture systems. The maximum difference between the reliability of a gait parameter in the marker-less and marker-based motion capture system was 0.3° , which was for the knee extension-flexion angle. Trunk flexion, pelvis abduction-adduction, and femur extension-flexion were the most reliable gait kinematic

parameters with the inter-trial reliability of less or equal to one degree in both marker-less and marker-based motion capture systems.

TABLE 4.10: The inter-trial reliability (*degrees*) of the kinematic gait parameters

	Parameter	Marker-less MoCap	Marker-based MoCap
Trunk	flexion	1.0	0.9
Pelvis	ab-adduction	0.7	0.7
	rotation	1.6	1.7
Hip	extension-flexion	1.3	1.4
	ab-adduction	1.2	1.1
Femur	extension-flexion	2.5	2.5
	ab-adduction	1.0	0.9
Knee	extension-flexion	2.2	1.9
	extension-flexion JB ¹	2.6	2.5

¹Joint center-Based.

4.3 Discussion

In this chapter, we aimed to evaluate the designed marker-less motion capture system on the ENSAM pose dataset. The marker-less motion capture system was designed for clinical gait analysis applications. Therefore, the evaluation metrics were not only joint position errors but also spatiotemporal and kinematic gait parameters.

4.3.1 Human pose estimation

The marker-less motion capture system consisted of four synchronized RGB cameras and deep learning-based human pose estimation methods. Comparing the performance of the human pose estimation method on the ENSAM test set, trained on the Human3.6M dataset, before and after finetuning on the ENSAM training set, showed a drastic improvement. For instance, for the ankle joint center, the 95% confidence intervals along the axes of the global coordinate system were within 68.5 mm. Finetuning the human pose estimation method on the ENSAM training set reduced these values to be less than or equal to 13.3 mm. These improvements may highlight the need for a well-adapted human pose dataset for training the human pose estimation methods which are the basis for the marker-less motion capture systems.

This incorporated human pose estimation method was selected because it achieved the most accurate results on the Human3.6M dataset. The achieved average Mean Per Joint Position Error (MPJPE) on the Human3.6M dataset across all actions was 17.7 mm, whereas for the single action of walking was 19.0 mm. However, the human pose estimation method on the ENSAM pose dataset achieved the average MPJPE of 15.3 mm. **Comparing the accuracy of the human pose estimation method on the ENSAM and Human3.6M dataset shows that the method's accuracy on the ENSAM pose dataset is around 20% better (19.0 mm – 15.3 mm = 3.7 mm).** Several reasons may explain the improvement of accuracy.

First, the ENSAM pose dataset was more accurate than the Human3.6M dataset. In the Human3.6M dataset, the markers were placed on the subjects' regular wearing. However, in the ENSAM pose dataset, the markers were placed directly on the subjects' anatomical landmarks by orthopedic surgeons. The bi-planar X-ray images from the subjects helped reconstruct the lower limb bony segments (Pelvis, femurs, tibias, and fibulas) and reflective markers. The joint

centers obtained from the reconstructed segments were registered to the marker-based motion capture environment using the reconstructed reflective markers. This technique reduced the errors of determining joint centers, which were mostly attributed to the marker misplacement on the anatomical landmarks. Therefore, in ENSAM pose dataset compared to the Human3.6M pose dataset, not only were the markers placed by professional individuals but also a technique helped reduce the errors attributed to the marker misplacement. Thus, the better accuracy of the ENSAM pose dataset, which was used to train and validate the human pose estimation method, may explain the human pose estimation method's improved accuracy.

Second, the training of the human pose estimation method is a crucial factor in determining its performance. As explained in section 4.1, the human pose estimation method consisted of two main parts – the 2D backbone human pose estimation method and volumetric triangulation. The 2D backbone was already pre-trained on the COCO dataset (Lin et al., 2014) and fine-tuned on the MPII and Human3.6M datasets. The whole human pose estimation method (2D backbone and volumetric triangulation) was then trained on the Human3.6M dataset. Then, we finetuned the trained network on the training set of the ENSAM pose dataset. In other words, we finetuned the human pose estimation method on the ENSAM pose dataset that was already trained on the Human3.6M dataset. Since finetuned models are generally more accurate than the trained models from scratch (Tajbakhsh et al., 2016), the training may be a reason for the improved accuracy of the human pose estimation method.

We stated that the 95th percentile for the shoulder position error of subject 04 was 240.4 *mm* that occurred in several frames in three out of ten walking trials. In these three walking trials, the subject was walking in the backward direction (opposite direction of walking toward frontal cameras). These high errors mostly (but not limited to) occurred because the human pose estimation method estimated the left shoulder instead of the right shoulder and vice versa. Nevertheless, these inconsistencies in the trajectory of a joint center could be easily identified by graphical visualization and filtered out.

The estimation accuracy of different joint centers, as shown in Table 4.6, is not the same. For instance, the ankles position error is 6.1 *mm*, whereas the trunk position error is 26.4 *mm*. Several factors may affect the estimation accuracy of different joint centers, including occlusion, the accuracy of joint centers' reference values.

Occlusion occurs when a joint is hidden (occluded) by another body segment or joint. As Figure 4.4 shows, in the last frame of the walking trial, the subject's left elbow and wrist are occluded by his body and cannot be seen by the lateral cameras while the joint centers can be in different positions. In the ENSAM pose dataset, in five out of ten walking trials, subjects were walking toward the frontal cameras of the marker-less motion capture system (forward direction), and in the other five walking trials, the subjects were walking in the opposite direction (backward direction). When subjects were walking in the forward direction, from the lateral cameras' view, the subjects' left elbow and wrist were mostly occluded by the subject's body and could be seen only by the frontal cameras. When a joint center is occluded, the accuracy of estimation decreases. The average joint position error, across all subjects' walking trials, of the right elbow and wrist, was 14.1 *mm* and 12.8 *mm*, whereas the left elbow and wrist's joint position error was 18.8 *mm* and 16.2 *mm*. Also, when the subjects were walking in the backward direction, from the lateral cameras' view, the subjects' right elbow and wrist were mostly occluded by the subject's body and could be seen only by frontal cameras. The right elbow and wrist's joint position error was 20.4 *mm* and 21.1 *mm*, whereas the left elbow and wrist's joint position error was 14.6 *mm* and 14.0 *mm*.

The accuracy of joint centers' reference values can be interpreted as the consistency of joint centers' reference values across different subjects. The trunk's reference values were retrieved from the marker placed on the tenth thoracic vertebrae (T10), or for the shoulder were retrieved from the markers placed on the acromion. The accuracy of these reference values was affected by the marker placement error. For example, the T10 marker might be placed slightly upper than a subject's T10 vertebrae, whereas, for another subject, the marker might be placed slightly lower than the T10 vertebrae. The human pose estimation method was trained and assessed using these reference values. Therefore, the less accurate the reference values were, the less accurate the estimations were. The trunk's and shoulders' joint position errors were the highest (refer to Table 4.6).

As we explained in the previous paragraphs, the lower limb joint centers, compared to the upper limbs, are less affected by marker placement error. Table 4.6 shows that the lower limb joint centers (ankles, knees, and pelvis) are among the most accurate joint centers. Nevertheless, the hip joint centers are less accurate than several upper limb joint centers (neck, head, wrists, and elbows) and less accurate than the pelvis. We believe that the pelvis joint center is estimated more accurately than the hip joint centers because the pelvis joint center's reference values are more accurate than the hip joint center's reference values. The hip joint centers are retrieved from the reconstructed femurs registered to the marker-based motion capture environment using the markers placed on the thighs, whereas the pelvis joint center is retrieved from the reconstructed pelvis registered to the marker-based system environment using the markers placed on the pelvis segment. The soft tissue artifact, which causes the displacement of markers with respect to the underlying bone, associated with the thigh is greater than the pelvis (Leardini et al., 2005). Since the soft-tissue artifact is a source of error that affects reference joint center values' accuracy, the pelvis' reference values are more accurate than the hips' reference values. Also, the pelvis center was defined as the midpoint (average) between the acetabulum's centers, whereas the hip joint center was defined as the femoral head center. Generally, an average is more robust than a point estimate.

4.3.2 Detection of gait events

The detection of gait events – heel strikes and toe-offs – is essential for determining the gait cycles required for gait analysis. Since we aimed to compare the marker-less with the marker-based motion capture system, a unique algorithm of gait events detection was implemented for marker-less and marker-based systems. To this end, we had modified the proposed method by (Zeni Jr, Richards, and Higginson, 2008) to be applicable for marker-less and marker-based motion capture systems because the marker-less motion capture system could not estimate the position of sacrum and heel markers. The ankle and pelvis joints were utilized instead of the heel and sacrum markers, respectively. Also, joint centers were low pass filtered using a zero-phase fourth-order Butterworth filter with the cutoff frequency 7 Hz. In the following paragraph, we illustrate that the applied modifications had a negligible effect on the performance of the gait event detection method.

The modified algorithm (implemented in this study) was compared with the original algorithm (Zeni Jr, Richards, and Higginson, 2008) for the marker-based motion capture system. The differences between the gait events detected by the original algorithm and the modified algorithm are shown in Table 4.11. The maximum difference between gait events was 2 frames (20 milliseconds). 97% of the differences were within 1 frame (10 milliseconds). The RMS difference was 0.78 frames (8 milliseconds). Therefore, the introduced differences because of the

modifications were negligible. And the modified method could be utilized to detect the gait events for marker-less and marker-based motion capture systems.

TABLE 4.11: The difference (*frames*) between the kinematic gait events detected based on the marker-based motion capture system utilizing the original algorithm (Zeni Jr, Richards, and Higginson, 2008) and the modified version

Parameter	MD ¹	SD ²	LLoA ³	ULoA ⁴	MxAD ⁵	RMSD ⁶
right heel strike	0.24	0.45	-0.64	1.13	2	0.51
left heel strike	0.30	0.49	-0.65	1.26	1	0.57
right toe off	-0.78	0.50	-1.77	1.20	2	0.93
left toe off	-0.81	0.54	-1.87	0.25	2	0.98

¹Mean Difference, ²Standard Deviation, ³Bland-Altman Lower Limit of Agreement, ⁴Bland-Altman Upper limit of agreement, ⁵Maximum Absolute Difference, ⁶Root Mean Square Difference.

(Kanko et al., 2020a) assessed the accuracy of the Theai3D marker-less motion capture system (introduced in section 1.3.2) for measuring the spatiotemporal parameters. In this study, the implemented algorithm for gait events detection (Zeni Jr, Richards, and Higginson, 2008) is similar to the one implemented in our study. The results of (Kanko et al., 2020a) showed that only about 80% of differences between the gait events, detected based on the Theai3D marker-less and a marker-based motion capture systems, were within two frames. However, in our study, 100% of the differences are within two frames (refer to Table 4.8). **This comparison may show that our designed marker-less motion capture system is more accurate than the Theai3D marker-less motion capture system (the state-of-the-art) in terms of gait event detection.**

4.3.3 Spatiotemporal gait parameters

We stated that gait analysis may be used for fall risk assessment or as a diagnostic and prognostic tool for several clinical conditions (refer to Section 1.1.3). For instance, the variability of **spatiotemporal gait parameters** could be utilized to identify future fallers in the older adult population. We measured the spatiotemporal gait parameters by the designed marker-less motion capture system and assessed its agreement with a marker-based motion capture system. In the following paragraph, we discuss that the marker-less system is accurate enough to measure the spatiotemporal gait parameters.

The **Minimum Detectable Change (MDC)** is defined as the minimum amount of change that is needed to identify a true performance change from a change due to the natural variability of the parameter or measurement error (Mohandas Nair, George Hornby, and Louis Behrman, 2012). The MDC of spatiotemporal parameters, based on marker-based motion capture systems, reported by several studies for different population studies, are shown in Table 4.12. The Bland-Altman lower and upper limits of agreement of all spatiotemporal parameters (refer to Table 4.8) were smaller than or equal to the minimum detectable changes. For instance, the lower and upper limits of agreement for gait speed were -0.007 m/s and 0.006 m/s , whereas the smallest minimum detectable change was 0.10 m/s . Therefore, **the marker-less motion capture system was accurate enough to estimate the spatiotemporal gait parameters, including gait speed, stride length, step length, step width, step time, stance time, swing time, and cadence.**

TABLE 4.12: Minimum detectable change (95% confidence interval) of spatiotemporal parameters, measured by marker-based motion capture systems, across different population studies.

Reference	Ref. ¹	Ref. ²	Ref. ³	Ref. ⁴	Ref. ⁵	Ref. ⁶	Ref. ⁷	Ref. ⁸
Population	Pop. ¹	Pop. ²	Pop. ³	Pop. ⁴	Pop. ⁵	Pop. ⁶	Pop. ⁷	Pop. ⁸
Gait speed (<i>m/s</i>)	0.16	0.12	0.10	–	0.17	0.17	–	0.13
Stride length (<i>cm</i>)	–	8	–	3	8	–	–	10
Step length (<i>cm</i>)	5	6	–	–	6	11	5	5
Step width (<i>cm</i>)	–	–	–	–	3	–	2	–
Step time (<i>sec</i>)	–	0.05	–	–	0.03	–	0.04	–
Stance time (<i>sec</i>)	–	0.07	–	–	–	–	0.03	0.06
Swing time (<i>sec</i>)	–	–	–	–	–	–	0.04	0.03
Cadence (<i>steps/sec</i>)	0.1	–	–	0.04	0.1	0.2	–	0.1

Pop.¹, MS population; Pop.², Chronic low back pain patients; Pop.³, hospitalized older fallers; Pop.⁴, Children; Pop.⁵, healthy subjects; Pop.⁶, population with incomplete spinal cord injury; Pop.⁷, older adults; Pop.⁸, people with Alzheimer’s disease.

Ref.¹, (Andreopoulou et al., 2019); Ref.², (Fernandes et al., 2015); Ref.³, (Hars, Herrmann, and Trombetti, 2013); Ref.⁴, (McSweeney, Reed, and Wearing, 2020); Ref.⁵, (Meldrum et al., 2014); Ref.⁶, (Mohandas Nair, George Hornby, and Louis Behrman, 2012); Ref.⁷, (Almarwani et al., 2016); Ref.⁸, (Wittwer et al., 2008).

4.3.4 Kinematic gait parameters

Assessing the agreement between the marker-less and marker-based motion capture systems in terms of joint kinematics (e.g., hip and knee joint angles) would be difficult because the marker-less motion capture system estimates only the joint center positions. For example, the marker-less system estimates the femoral head (hip joint center) and the center of the lateral and medial femur condyles (knee joint center), but these two centers form only one axis of the femur coordinate system. We attempted to address this issue and assess the agreement between the marker-less and marker-based motion capture system by two different approaches.

In the first approach, the joint kinematics were computed based on joint center positions – joint center-based knee extension-flexion, femur abduction-adduction, femur extension-flexion, and trunk flexion. The maximum RMS difference among these four parameters was 2.5° for the Joint center-Based (JB) knee extension-flexion angle, which had the highest motion range. Also, for the JB knee extension-flexion, the Bland-Altman lower and upper limits of agreement were -4.8° and 4.8°. (McGinley et al., 2009) state that the errors less than 5° are “likely to be regarded as reasonable in gait analysis errors.” Therefore, if we assume the 5° as the acceptable limits, **we could consider the marker-less motion capture system accurate enough to estimate the joint center-based kinematic gait parameters, including JB knee extension-flexion, femur abduction-adduction, femur extension-flexion, and trunk flexion.**

In the second approach, the bony segments’ coordinate system was computed, partially based on the marker-less motion capture data. The marker-less motion capture system could retrieve only one axis of the bony segments’ coordinate system. The second axis was retrieved from the axis of the bony segments’ coordinate system, averaged across a control group, obtained by a marker-based motion capture system. The Bland-Altman lower and upper limits of agreement for the coordinate system-based kinematic gait parameters, including pelvis abduction-adduction, rotation, hip abduction-adduction, extension-flexion, and knee extension-flexion, were not consistent with the limit of 5°. This inconsistency generally happened because

using an average axis for the formation of the bony segments' coordinate system across all subjects led to a bias between the marker-less and marker-based motion capture systems. Even though **there was no adequate agreement between the marker-less and marker-based motion capture systems for the coordinate systems-based kinematic gait parameters**, the inter-trial reliability (refer to Table 4.10) showed that **the marker-less motion capture system was almost as reliable as the marker-based motion capture system**.

4.3.5 Limitations

In this study, a limitation was that the subjects' bounding boxes were determined using the marker-based motion capture system. Several approaches could help determine the bounding boxes. For instance, for a captured sequence, in the sequence's first frame, we could manually crop the subject's bounding boxes; then, the remaining bounding boxes would be retrieved iteratively from the estimated joints in the previous frames. Also, the Mask R-CNN (He et al., 2017), which is a convolutional neural network used for segmentation, could help retrieve the subject's bounding boxes across a sequence's frames.

A general limitation of the designed marker-less motion capture system is that it can only estimate the joint centers. Subsequently, several kinematic gait parameters (e.g., hip extension-flexion) cannot be directly calculated. Further studies would seem required to develop this system to be able to measure all kinematic gait parameters.

Another limitation of this study was that the in-vivo data acquisitions were performed in a single laboratory environment. Subsequently, the training and test of the marker-less motion capture system were conducted using those data. Therefore, the performance of the system was not assessed across different environments. Nonetheless, a single gait analysis session was held for one of the subjects with XLH at the Kremlin-Bicêtre hospital, Paris, using the marker-less motion capture system. Still, the measured spatiotemporal and kinematic gait parameters are to be carefully examined.

Also, even though compared to the state-of-the-art marker-less motion capture systems (e.g., Theai3D marker-less system, refer to Section 1.3.2), this system uses a fewer number of RGB cameras, four cameras can still be regarded as a limitation. A fewer number of cameras make the system easier-to-use. In future studies, various pose estimation methods with a fewer number of cameras should be analyzed.

4.4 Conclusion

In this section, we assessed the performance of the designed marker-less motion capture system against a marker-based motion capture system in terms of the joint center error, detection of gait events, spatiotemporal and kinematic gait parameters.

In chapter 1, we set the design criteria for the marker-less motion capture system. Herein, we review them to check whether the design criteria are met. The designed marker-less motion capture is of low-cost. The total price of equipment, including four GoPro Hero 7 Black cameras (4×350 €), the Smart Remote (100 €), calibration device (< 200 €), synchronization device (< 100 €), tripods and mounting devices (< 100 €), processing unit which the main component is GPU (< 2500 €), is less than 5000 €.

The designed marker-less motion capture system is accurate enough in terms of spatiotemporal gait parameters – gait speed, stride length, step length, step width, step time, stance time, swing time, and cadence. In terms of kinematic gait parameters, the evaluation is difficult because the designed marker-less motion capture system can only estimate the joint centers and cannot estimate the bony segments' coordinate system. To resolve this issue, two initiatives were proposed – the joint center-based parameters and formation of coordinate systems partly using a priori knowledge. The error (Bland-Altman lower and upper limits of agreement) of the joint center-based kinematic gait parameters is less than 5 degree and can be regarded as accurate enough for clinical applications. The error of the coordinate system-based parameters was not less than 5 degrees, whereas the reliability of all kinematic gait parameters was consistent with marker-based motion capture systems.

In terms of the laboratory environment, since we aimed to compare the designed marker-less motion capture system against a marker-based motion capture system, the experiment was performed in a laboratory environment. However, the RGB cameras of the designed marker-less system, compared to the cameras of marker-based systems, are not sensitive to ambient light or light reflections. Therefore, we believe that the designed marker-less system is not limited to laboratory environments.

The designed motion capture system is marker-less which helps reduce the patient preparatory time. The designed system is automatic, and there is no need for manual intervention. The processing time varies depending on the processing unit. Nevertheless, for a captured sequence of 300 frames, the processing time on average is less than 5 – 10 minutes.

Therefore, the designed marker-less motion capture system may hold a great promise to be utilized to measure gait parameters in clinical applications.

General Conclusion and Perspectives

Walking ability is of invaluable importance in life quality. Falls are a threat to the health and independence of the older adult population. Gait analysis, ability evaluation of walking can help to identify future fallers in the more aging adult population and apply preventive measures to reduce the risk of falling. Gait analysis also can be used as a diagnostic and prognostic tool for several clinical conditions (e.g., cerebral palsy). The current gold standard instrumentation for gait analysis is marker-based motion capture systems. However, limitations of these systems, such as equipment costs, restrain the wide-spread use of gait analysis in clinics. The objective of this study was to develop an accurate, easy-to-use, and cost-effective motion capture system that may help to wide-spread use of gait analysis for clinical applications.

Among the existing methods, we focused on the human pose estimation methods based on RGB images to develop the motion capture system. These methods had achieved promising results on the publicly available datasets (e.g., Human3.6M) while their performance for measuring gait parameters was unknown. We proposed a marker-less motion capture system based on four RGB cameras and state-of-the-art deep learning-based human pose estimation methods. Even though most of the marker-less systems were based on eight cameras, we selected four number of cameras because to make the system easier-to-use. Also, since the performance of the deep learning-based human pose estimation methods depends on the training dataset, a dedicated dataset (ENSAM dataset), well-adapted for clinical gait study, was collected.

Publicly available datasets (e.g., Human3.6M dataset) may be less efficient for gait study, particularly for clinical applications, because of the limited number of subjects (only seven subjects) and their homogeneity (all asymptomatic adults), and the errors introduced by marker placement on subjects' regular clothing. ENSAM dataset contained the walking trials of thirty-one asymptomatic and pathologic subjects from a wide range of age. Particular care was given to the accuracy of data collection. Not only the accurate placement of markers on anatomical landmarks were performed by orthopedic surgeons, but also a medical imaging system (EOS) was used for accurate registration between joint centers and external reflective markers.

The ENSAM pose dataset was utilized for training and evaluating the proposed marker-less motion capture system. Because of the size of the dataset, 310 walking trials performed by thirty-one subjects corresponding to more than 93,000 multi-view frames, the dataset could be split into training (19 subjects) and test sets (12 subjects). The human pose estimation method was trained on the Human3.6M dataset, then finetuned on the ENSAM training set. The performance of this system was evaluated on the test set in terms of joint position errors following the metric introduced by the Human3.6M dataset. Also, following the objective of this study, the estimation accuracy of spatiotemporal and kinematic gait parameters was evaluated.

Comparing the performance of the human pose estimation method on the ENSAM test set, trained on the Human3.6M dataset, before and after finetuning on the ENSAM training set, showed a drastic improvement – 95% confidence interval for hip joint center along Z-axis reduced from 106.3 mm to 19.8 mm. Also, the proposed marker-less system achieved state-of-the-art results in terms of joint position error for walking (mean per joint position error reduced

from 19.0 mm to 15.3 mm). In terms of gait parameters, the achieved result showed that our system was more accurate than the Theai3D system, marker-less motion capture system to detect gait events. Concerning the spatiotemporal gait parameters, the results showed that the Bland-Altman lower and upper limits of agreement were smaller than the minimum detectable changes (the minimum change needed to identify a true performance change) reported by different studies. Therefore, the designed system is accurate enough to measure spatiotemporal gait parameters.

Regarding the kinematic gait parameters, the Bland-Altman upper and lower limits of agreement for joint center-based parameters were smaller than 5 degrees, demonstrating that the errors are reasonable for clinical applications. However, since the marker-less motion capture system could only estimate the joint centers, several kinematic gait parameters such as hip extension-flexion or ab-adduction could not be measured directly. Further studies are necessary to advance the marker-less system to be able to measure those kinematic gait parameters, and work is in progress in the Institute for that aim.

Another limitation is that, even though this system only uses to pairs of RGB cameras, four cameras can still be regarded as a limitation since a fewer number of cameras would make the system easier-to-use. Also, even though this study achieved promising results for clinical gait analysis, the results were preliminary. The number of subjects is still limited and not representative of the range of subjects with or without musculoskeletal troubles (age ≤ 44 years, particularly lack of older adults). However, data collection protocol is now well defined and further data collection can easily be performed. Improving the dataset may increase the accuracy of the system, and allow progressing towards reduction of the number of cameras.

We already attempted to use the marker-less system in clinics. A gait analysis session was conducted at the Kremlin-Bicêtre hospital, Paris (Figure 1), and this preliminary session demonstrated the feasibility of this new user friendly approach. Work is still in progress to carefully examine the reliability of spatiotemporal and kinematic gait parameters. This system is operational to be used in clinical research and may pave the way for using gait analysis as a clinical routine, thus, improving the care for the society.



FIGURE 1: Gait analysis session conducted at the Kremlin-Bicêtre hospital, Paris, using the designed marker-less motion capture system. Four images are captured by the four RGB cameras of the marker-less system.

Appendix A

Pinhole camera model

The simplest camera model is the pinhole camera model (Bradski and Kaehler, 2008a). The pinhole camera model works flawlessly when the camera lenses are perfect. However, in reality, there is no perfect lens. Radial and tangential distortions, which may appear because of the lens manufacturing and assembly process (Bradski and Kaehler, 2008b), should be modeled when high accuracy is expected. According to the pinhole camera model and distortion model (Bouguet, 2005), given $Q = [X, Y, Z]^T$ in the global coordinate system, Equation A.1 transforms the Q to the camera coordinate system using the extrinsic parameters.

$$Q_c = R \times Q + T \rightarrow \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + T \quad (\text{A.1})$$

where R and T represent the rotation matrix and translation vector, respectively. These parameters are called **extrinsic parameters**.

Equation A.2 applies the distortion model:

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} X_c/Z_c \\ Y_c/Z_c \end{bmatrix} \\ \begin{bmatrix} x_d \\ y_d \end{bmatrix} &= \begin{bmatrix} (1 + k_1 r^2 + k_2 r^4 + k_3 r^6)x + 2p_1 xy + p_2(r^2 + 2x^2) \\ (1 + k_1 r^2 + k_2 r^4 + k_3 r^6)y + p_1(r^2 + 2y^2) + 2p_2 xy \end{bmatrix}; \quad r^2 = x^2 + y^2 \end{aligned} \quad (\text{A.2})$$

where k_1, k_2, k_3 represent the radial distortion parameters, and p_1, p_2 represent the tangential distortion parameters. These parameters are a part of the **intrinsic parameters**.

Finally, Equation A.3 computes the 2D coordinates $q = [x_p, y_p]^T$.

$$\begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} = K \begin{bmatrix} x_d \\ y_d \\ 1 \end{bmatrix}; \quad K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.3})$$

where f_x, f_y represent the focal lengths, c_x, c_y represent the principal points, and s represents the skew coefficient. These parameters are a part of the **intrinsic parameters**.

In a preliminary test, we evaluated the accuracy of the intrinsic and extrinsic calibration. 37 different images¹ were captured from a printed checkerboard. The printed checkerboard consisted of 14 rows and 20 columns of 20 mm-width-checkers. Next, the corners on the checkerboard were extracted as the control points, and the camera model parameters were estimated. The Stereo Camera Calibrator App estimated each of two cameras' intrinsic parameters and extrinsic parameters. The extrinsic parameters determine the relative position and orientation of two cameras (frontal cameras) with respect to each other.

¹Frontal cameras are two cameras. Capturing 37 images by frontal cameras means that each camera captures 37 images.

Then, another printed checkerboard, which consisted of 4 rows and 5 columns of 70 mm-width-checkers, was placed at 35 (5 columns, 7 rows) different places (as shown in Figure 1), covering the entire capture space. The checkers' 12 corners were extracted, and the 3D positions of corners were constructed using the linear triangulation method. The 17 checkers' sides width were computed using the 3D positions of corners. Table A.1 shows the obtained results.

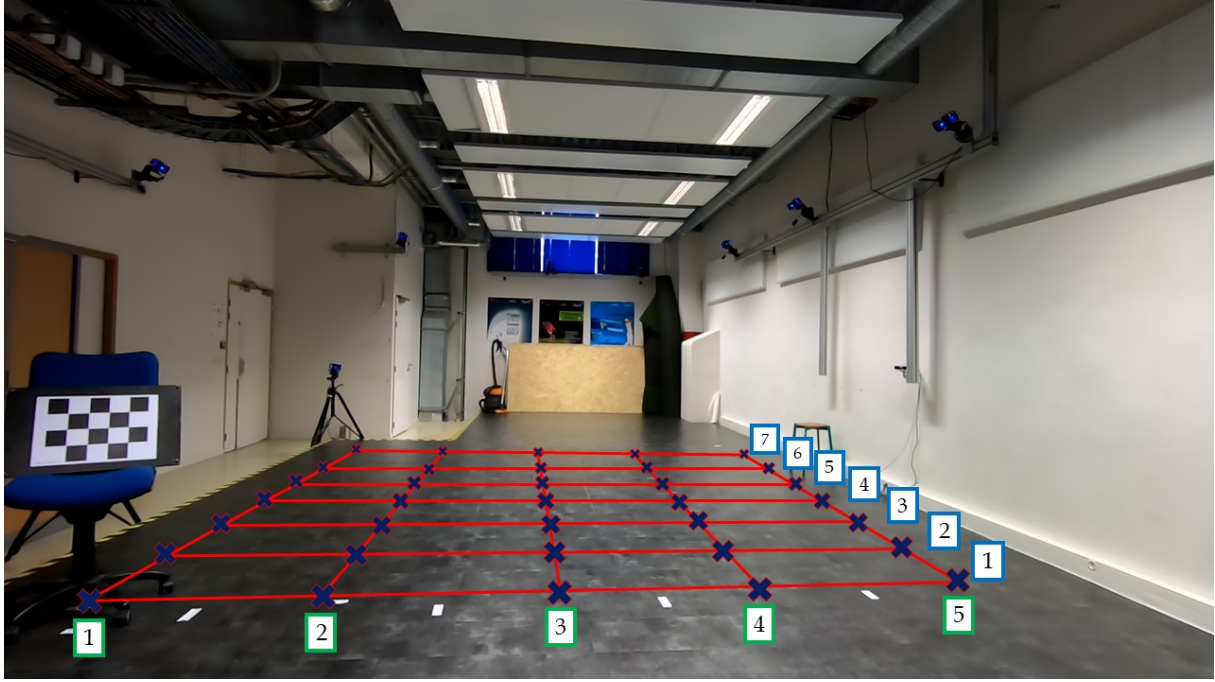


FIGURE 1: The checkerboard placement at 35 different positions for measuring the checker's sides width using the frontal cameras.

TABLE A.1: RMS error (*millimeter*) of measuring the checkerboard's 17 checkers' sides widths placed at 35 different places covering the entire capture space. The 5 columns and 7 rows denote the displayed columns and rows in Figure 1.

	Column #1	Column #2	Column #3	Column #4	Column #5
Row #1	1.4	0.6	0.2	0.3	0.9
Row #2	1.3	0.5	0.3	0.4	0.5
Row #3	0.8	0.5	0.3	0.5	0.7
Row #4	1.6	0.6	0.4	0.8	0.5
Row #5	1.4	0.7	0.6	0.6	1.2
Row #6	1.1	1.0	0.5	0.5	0.9
Row #7	2.2	1.0	0.4	1.1	1.3

The reported results in Table A.1 shows that the maximum RMS error was 2.2 *mm*. 85% of the RMS errors in the second to fourth columns were below 1 *mm*. The remaining 15% were below or equal to 2.2 *mm* mostly being at the farthest position with respect to the frontal cameras. Generally, the RMS errors in the first and fifth columns, which were the lateral sides of the capture space, were higher than the second to fourth columns. We believe that there were two underlying reasons for these errors. First, the accuracy of corner extraction degrades while the checkerboard was at the farthest position. Second, on the lateral sides, the lens distortion effect had a higher impact.

Appendix B

Marker-set

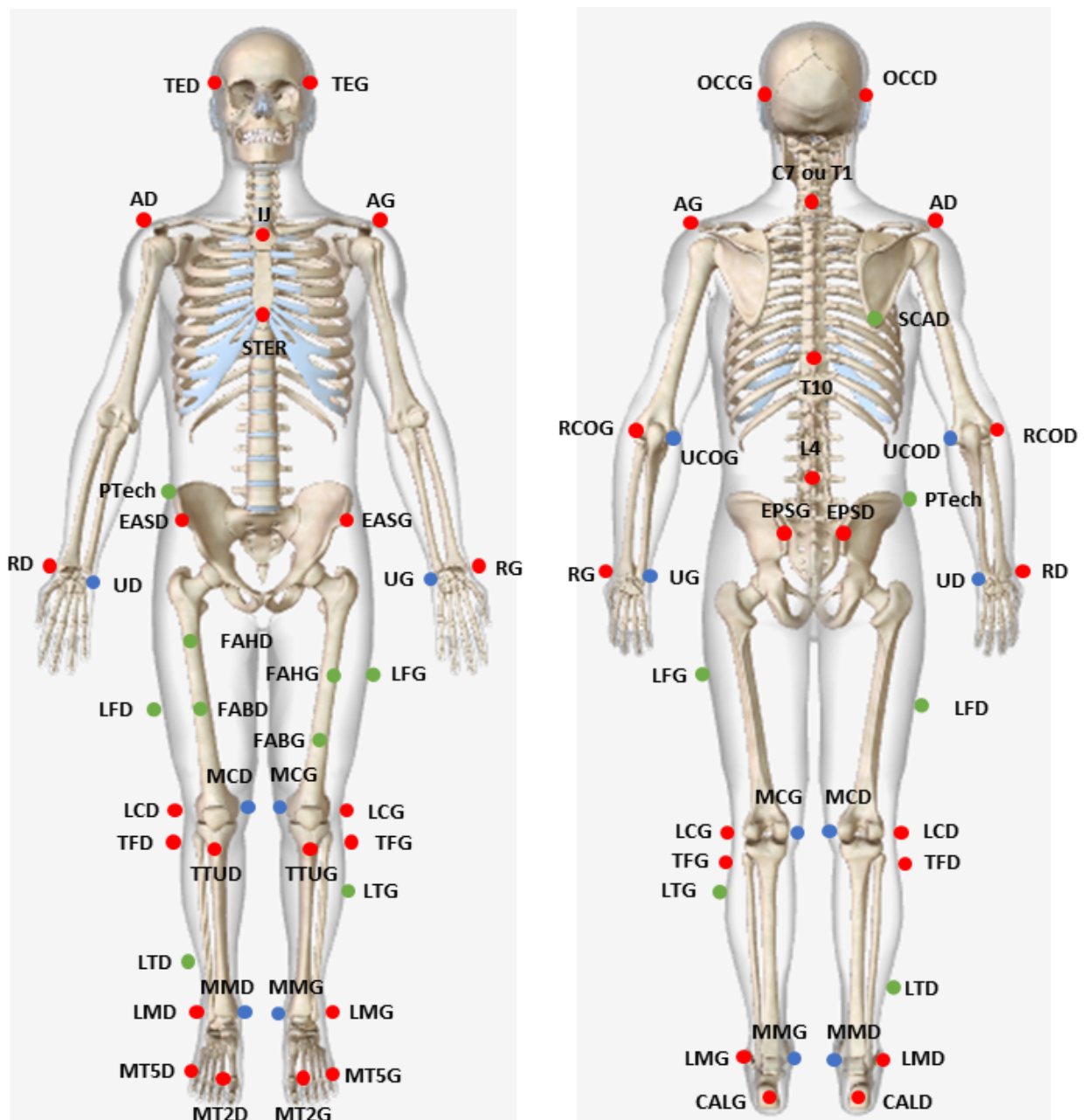


FIGURE 1: The marker-set used for the marker placement on the subjects' anatomical landmarks.

Red, green, and blue markers represent anatomical, technical, and static markers. The description of anatomical markers and static is given in Table B.1. The technical markers are placed approximately on the body segments, as shown in Figure 1.

TABLE B.1: Anatomical landmarks

Lower limbs	
MT2D & MT2G	Right and left second metatarsal head
MT5D & MT5G	Right and left fifth metatarsal head
CALD & CALG	Right and left calcaneus
LMD & LMG	Right and left lateral malleolus
MMD & MMG	Right and left medial malleolus
TTUD & TTUG	Right and left tibial tuberosity
TFD & TFG	Right and left head of fibula
LCD & LCG	Right and left lateral condyle
MCD & MCG	Right and left medial condyle
EPSD & EPSG	Right and left posterior superior iliac spine
EASD & EASG	Right and left anterior superior iliac spine
Axial structure	
C7/T1	Seventh cervical vertebra / First thoracic vertebrae
T10	Tenth thoracic vertebrae
L4	Fourth lumbar vertebra
STER	Tip of the sternum
IJ	Incisure jugulaire
Upper limbs	
TED & TEG	Right and left head temporal bone
OCCD & OCCG	Right and left head occipital bone
AD & AG	Right and left acromion
RCOD & RCOG	Right and left proximal end of radius
UCOD & UCOG	Right and left proximal end of ulnar
RD & RG	Right and left distal end of radius
UD & UG	Right and left distal end of ulnar

Bibliography

- Agarwal, Ankur and Bill Triggs (2005). "Recovering 3D human pose from monocular images". In: *IEEE transactions on pattern analysis and machine intelligence* 28.1, pp. 44–58.
- Almarwani, Maha et al. (2016). "The test-retest reliability and minimal detectable change of spatial and temporal gait variability during usual over-ground walking for younger and older adults". In: *Gait & posture* 44, pp. 94–99.
- Amin, Sikandar et al. (2013). "Multi-view pictorial structures for 3d human pose estimation." In: *Bmvc*. Vol. 1, p. 2.
- Andreopoulou, Georgia et al. (2019). "Test-retest reliability and minimal detectable change of ankle kinematics and spatiotemporal parameters in MS population". In: *Gait & Posture* 74, pp. 218–222.
- Andriluka, Mykhaylo et al. (2014). "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Anguelov, Dragomir et al. (2005). "SCAPE: shape completion and animation of people". In: *ACM SIGGRAPH 2005 Papers*, pp. 408–416.
- Arnab, Anurag, Carl Doersch, and Andrew Zisserman (2019). "Exploiting temporal context for 3D human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404.
- Attias, Michael et al. (2019). "Kinematics can help to discriminate the implication of iliopsoas, hamstring and gastrocnemius contractures to a knee flexion gait pattern". In: *Gait & Posture* 68, pp. 415–422.
- Baker, Richard (2006). "Gait analysis methods in rehabilitation". In: *Journal of neuroengineering and rehabilitation* 3.1, p. 4.
- Baker, RJ et al. (2018). "The conventional gait model-success and limitations". In: *Handbook of human motion*, pp. 489–508.
- Becker, Linda and Ph Russ (2015). "Evaluation of joint angle accuracy using markerless silhouette based tracking and hybrid tracking against traditional marker tracking". In: *Poster für Masterarbeit bei Simi Reality Motion Systems GmbH und der Otto-von-Guericke-Universität Magdeburg*.
- Belagiannis, Vasileios and Andrew Zisserman (2017). "Recurrent human pose estimation". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, pp. 468–475.
- Belagiannis, Vasileios et al. (2014). "3D pictorial structures for multiple human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1669–1676.
- Bell, Alexander L, Douglas R Pedersen, and Richard A Brand (1990). "A comparison of the accuracy of several hip center location prediction methods". In: *Journal of biomechanics* 23.6, pp. 617–621.
- Belongie, Serge, Jitendra Malik, and Jan Puzicha (2001). "Shape context: A new descriptor for shape matching and object recognition". In: *Advances in neural information processing systems*, pp. 831–837.
- Bem, Rodrigo de et al. (2018). "Deep fully-connected part-based models for human pose estimation". In: *Asian Conference on Machine Learning*, pp. 327–342.

- Benedetti, Maria Grazia et al. (2017). "SIAMOC position paper on gait analysis in clinical practice: General requirements, methods and appropriateness. Results of an Italian consensus conference". In: *Gait & posture* 58, pp. 252–260.
- Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.
- Bi, Xiaojun and Xuelian Zou (2019). "Human Pose Estimation Based on Improved Hourglass Networks". In: *2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019)*. Atlantis Press.
- Biswas, Sandika et al. (2019). "Lifting 2d Human Pose to 3d: A Weakly Supervised Approach". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–9.
- Bonnechère, B. et al. (2014). "Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry". In: *Gait Posture* 39.1, pp. 593–598. ISSN: 0966-6362. DOI: <https://doi.org/10.1016/j.gaitpost.2013.09.018>. URL: <http://www.sciencedirect.com/science/article/pii/S0966636213006310>.
- Bouguet, JY (2005). "Camera calibration toolbox for MATLAB. California Institute of Technology". In: *Computational Vision at CALTECH*.
- Bradski, Gary and Adrian Kaehler (2008a). "Learning OpenCV: Computer vision with the OpenCV library". In: "O'Reilly Media, Inc.", p. 370.
- (2008b). "Learning OpenCV: Computer vision with the OpenCV library". In: "O'Reilly Media, Inc.", p. 375.
- Bulat, Adrian and Georgios Tzimiropoulos (2016). "Human Pose Estimation via Convolutional Part Heatmap Regression". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 717–732. ISBN: 978-3-319-46478-7.
- (2017). "Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment With Limited Resources". In: *The IEEE International Conference on Computer Vision (ICCV)*.
- Caldas, Rafael et al. (2017). "A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms". In: *Gait & posture* 57, pp. 204–210.
- Cao, Zhe et al. (2018). "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields". In: *arXiv preprint arXiv:1812.08008*.
- Cappozzo, Aurelio (1984). "Gait analysis methodology". In: *Human Movement Science* 3.1-2, pp. 27–50.
- Cappozzo, Aurelio et al. (2005). "Human movement analysis using stereophotogrammetry: Part 1: theoretical background". In: *Gait & posture* 21.2, pp. 186–196.
- Ceseracciu, Elena, Zimi Sawacha, and Claudio Cobelli (2014). "Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept". In: *PloS one* 9.3, e87640.
- Chaibi, Y et al. (2012). "Fast 3D reconstruction of the lower limb using a parametric model and statistical inferences and clinical measurements calculation from biplanar X-rays". In: *Computer methods in biomechanics and biomedical engineering* 15.5, pp. 457–466.
- Chalmers, Eric et al. (2014). "Inertial sensing algorithms for long-term foot angle monitoring for assessment of idiopathic toe-walking". In: *Gait & posture* 39.1, pp. 485–489.
- Chen, Ching-Hang et al. (2019a). "Unsupervised 3d pose estimation with geometric self-supervision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5714–5724.
- Chen, Xipeng et al. (2019b). "Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10895–10904.

- Chen, Yu et al. (2017). "Adversarial posenet: A structure-aware convolutional network for human pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1212–1221.
- Chiari, Lorenzo et al. (2005). "Human movement analysis using stereophotogrammetry: Part 2: Instrumental errors". In: *Gait & posture* 21.2, pp. 197–211.
- Chou, Chia-Jung, Jui-Ting Chien, and Hwann-Tzong Chen (2018). "Self adversarial training for human pose estimation". In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 17–30.
- Chu, Xiao et al. (2017). "Multi-Context Attention for Human Pose Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Clark, Ross A et al. (2019). "Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives". In: *Gait & posture* 68, pp. 193–200.
- CMU Graphics Lab Motion Capture Database. URL: <http://mocap.cs.cmu.edu/>. (accessed: 24.09.2019).
- Coutts, Fiona (1999). "Gait analysis in the therapeutic environment". In: *Manual therapy* 4.1, pp. 2–10.
- Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE, pp. 886–893.
- Data Curation. online, accessed 28 October 2020. URL: <https://guide.dhcuration.org/contents/data-representation/>.
- Davis, Roy B et al. (1991). "A gait analysis data collection and reduction technique". In: Della Croce, Ugo, Aurelio Cappozzo, and D Casey Kerrigan (1999). "Pelvis and lower limb anatomical landmark calibration precision and its propagation to bone geometry and joint angles". In: *Medical & biological engineering & computing* 37.2, pp. 155–161.
- Della Croce, Ugo et al. (2005). "Human movement analysis using stereophotogrammetry: Part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics". In: *Gait & posture* 21.2, pp. 226–237.
- DeLuca Peter; Davis, Roy; Öunpuu Sylvia; Rose Sally; Sirkin Robert (1997). "Alterations in Surgical Decision Making in Patients with Cerebral Palsy Based on Three-Dimensional Gait Analysis". In: *Journal of Pediatric Orthopaedics* 17 (5), pp. 608–614. URL: https://journals.lww.com/pedorthopaedics/Abstract/1997/09000/Alterations_in_Surgical_Decision_Making_in.7.aspx.
- Dimitrijevic, Miodrag, Vincent Lepetit, and Pascal Fua (2006). "Human body pose detection using Bayesian spatio-temporal templates". In: *Computer vision and image understanding* 104.2–3, pp. 127–139.
- Dorschky, Eva et al. (2019). "Estimation of gait kinematics and kinetics from inertial sensor data using optimal control of musculoskeletal models". In: *Journal of biomechanics* 95, p. 109278.
- Elhayek, Ahmed et al. (2016). "MARCONI—ConvNet-based MARKer-less motion capture in outdoor and indoor scenes". In: *IEEE transactions on pattern analysis and machine intelligence* 39.3, pp. 501–514.
- Felzenszwalb, Pedro F and Daniel P Huttenlocher (2005). "Pictorial structures for object recognition". In: *International journal of computer vision* 61.1, pp. 55–79.
- Fernandes, Rita et al. (2015). "Test–retest reliability and minimal detectable change of three-dimensional gait analysis in chronic low back pain patients". In: *Gait & posture* 42.4, pp. 491–497.
- Ferrarin, M et al. (2015). "Does gait analysis change clinical decision-making in poststroke patients? Results from a pragmatic prospective observational study". In: *Eur J Phys Rehabil Med* 51.2, pp. 171–84.

- Freifeld, Oren et al. (2010). "Contour people: A parameterized model of 2D articulated human shape". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 639–646.
- Frühschütz, Hannes et al. (2017). "Evaluation of Silhouette-based Markerless Tracking for Kinematics in Sport". In: *ISBS Proceedings Archive* 35.1, p. 231.
- Fukushima, Kunihiko and Sei Miyake (1982). "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, pp. 267–285.
- Fuller, David A et al. (2002). "The impact of instrumented gait analysis on surgical planning: treatment of spastic equinovarus deformity of the foot and ankle". In: *Foot & ankle international* 23.8, pp. 738–743.
- Furtado, Joshua S et al. (2019). "Comparative analysis of optitrack motion capture systems". In: *Advances in Motion Sensing and Control for Robotic Applications*. Springer, pp. 15–31.
- Gong, Wenjuan et al. (2016). "Human Pose Estimation from Monocular Images: A Comprehensive Survey". In: *Sensors* 16.12. ISSN: 1424-8220. DOI: [10.3390/s16121966](https://doi.org/10.3390/s16121966). URL: <http://www.mdpi.com/1424-8220/16/12/1966>.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.
- Goulermas, John Y et al. (2008). "An instance-based algorithm with auxiliary similarity information for the estimation of gait kinematics from wearable sensors". In: *IEEE transactions on neural networks* 19.9, pp. 1574–1582.
- Gu, Jiuxiang et al. (2018). "Recent advances in convolutional neural networks". In: *Pattern Recognition* 77, pp. 354–377.
- Guler, Riza Alp and Iasonas Kokkinos (2019). "Holopose: Holistic 3d human reconstruction in-the-wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10884–10894.
- Gundavarapu, Nitesh B et al. (2019). "Structured Aleatoric Uncertainty in Human Pose Estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 50–53.
- Hamacher, D et al. (2011). "Kinematic measures for assessing gait stability in elderly individuals: a systematic review". In: *Journal of The Royal Society Interface* 8.65, pp. 1682–1698.
- Hars, Mélanie, François R Herrmann, and Andrea Trombetti (2013). "Reliability and minimal detectable change of gait variables in community-dwelling and hospitalized older fallers". In: *Gait & Posture* 38.4, pp. 1010–1014.
- Harsted, Steen et al. (2019). "Concurrent validity of lower extremity kinematics and jump characteristics captured in pre-school children by a markerless 3D motion capture system". In: *Chiropractic & manual therapies* 27.1, p. 39.
- Hausdorff, Jeffrey M, Dean A Rios, and Helen K Edelberg (2001). "s". In: *Archives of physical medicine and rehabilitation* 82.8, pp. 1050–1056.
- He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- He, Kaiming et al. (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.
- Horn, Berthold KP and Brian G Schunck (1981). "Determining optical flow". In: *Techniques and Applications of Image Understanding*. Vol. 281. International Society for Optics and Photonics, pp. 319–331.
- Hu, Peiyun and Deva Ramanan (2016). "Bottom-Up and Top-Down Reasoning With Hierarchical Rectified Gaussians". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Huang, Yinghao et al. (2017). "Towards accurate marker-less human shape and pose estimation over time". In: *2017 international conference on 3D vision (3DV)*. IEEE, pp. 421–430.

- Human3.6M Benchmark*. online, accessed 22 September 2020. URL: <https://paperswithcode.com/sota/3d-human-pose-estimation-on-human36m>.
- Immonen, Milla (2020). "Risk factors for falls and technologies for fall risk assessment in older adults". In:
- Insafutdinov, Eldar et al. (2016). "DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 34–50. ISBN: 978-3-319-46466-4.
- Ionescu, C. et al. (2014). "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7, pp. 1325–1339. ISSN: 0162-8828. DOI: [10.1109/TPAMI.2013.248](https://doi.org/10.1109/TPAMI.2013.248).
- Iskakov, Karim et al. (2019). "Learnable triangulation of human pose". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7718–7727.
- Jiang, Mengxi et al. (2019). "Reweighted sparse representation with residual compensation for 3D human pose estimation from a single RGB image". In: *Neurocomputing* 358, pp. 332 – 343. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2019.05.034>. URL: <http://www.sciencedirect.com/science/article/pii/S0925231219307179>.
- Johnson, Sam and Mark Everingham (2010). "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." In: *bmvc*. Vol. 2. 4. Citeseer, p. 5.
- (2011). "Learning effective human pose estimation from inaccurate annotation". In: *CVPR 2011*. IEEE, pp. 1465–1472.
- Joo, Hanbyul et al. (2015). "Panoptic studio: A massively multiview system for social motion capture". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3334–3342.
- Joo, Hanbyul et al. (2017). "Panoptic Studio: A Massively Multiview System for Social Interaction Capture". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ju, Shanon X, Michael J Black, and Yaser Yacoob (1996). "Cardboard people: A parameterized model of articulated image motion". In: *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*. IEEE, pp. 38–44.
- Kanazawa, Angjoo et al. (2018). "End-to-end recovery of human shape and pose". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131.
- Kanko, Robert et al. (2020a). "Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system". In:
- Kanko, Robert M et al. (2020b). "Inter-session repeatability of Theia3D markerless motion capture gait kinematics". In: *BioRxiv*.
- Kawakami, Hideo et al. (2005). "Gait analysis system for assessment of dynamic loading axis of the knee". In: *Gait & posture* 21.1, pp. 125–130.
- Kawana, Yuki et al. (2018). "Ensemble convolutional neural networks for pose estimation". In: *Computer Vision and Image Understanding* 169, pp. 62–74.
- Ke, Lipeng et al. (2018). "Multi-scale structure-aware network for human pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 713–728.
- Kernozek, T. W. and J. D. Willson (2018). URL: <https://musculoskeletalkey.com/gait/>.
- Kocabas, Muhammed, Salih Karagoz, and Emre Akbas (2019). "Self-supervised learning of 3d human pose using multi-view geometry". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1086.
- Kovalenko, Onorina et al. (2019). "Structure from Articulated Motion: An Accurate and Stable Monocular 3D Reconstruction Approach without Training Data". In: *arXiv preprint arXiv:1905.04789*.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.
- Leardini, Alberto et al. (1999). "Validation of a functional method for the estimation of hip joint centre location". In: *Journal of biomechanics* 32.1, pp. 99–103.

- Leardini, Alberto et al. (2005). "Human movement analysis using stereophotogrammetry: Part 3. Soft tissue artifact assessment and compensation". In: *Gait & posture* 21.2, pp. 212–225.
- Lee, Kyoungoh, Inwoong Lee, and Sanghoon Lee (2018). "Propagating lstm: 3d pose estimation based on joint interdependency". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–135.
- Liang, X. et al. (2018). "Look into Person: Joint Body Parsing and Pose Estimation Network and a New Benchmark". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. ISSN: 0162-8828. DOI: [10.1109/TPAMI.2018.2820063](https://doi.org/10.1109/TPAMI.2018.2820063).
- Lin, Tsung-Yi et al. (2014). "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer, pp. 740–755.
- Liu, Ding et al. (2019a). "Improving 3D human pose estimation via 3D part affinity fields". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1004–1013.
- Liu, Jian, Naveed Akhtar, and Ajmal Mian (2019). "Temporally Coherent Full 3D Mesh Human Pose Recovery from Monocular Video". In: *arXiv preprint arXiv:1906.00161*.
- Liu, Jun et al. (2019b). "Feature boosting network for 3D pose estimation". In: *IEEE transactions on pattern analysis and machine intelligence*.
- Liu, Wentao et al. (2018). "A cascaded inception of inception network with attention modulated feature fusion for human pose estimation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Loper, Matthew et al. (Oct. 2015). "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graph.* 34.6. ISSN: 0730-0301. DOI: [10.1145/2816795.2818013](https://doi.org/10.1145/2816795.2818013). URL: <https://doi.org/10.1145/2816795.2818013>.
- Lord, S.R., C. Sherrington, and H.B. Menz (2001). *Falls in Older People: Risk Factors and Strategies for Prevention*. Cambridge University Press. ISBN: 9780521589642. URL: <https://books.google.fr/books?id=nWlCqLGiktkC>.
- Lowe, David G (1999). "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee, pp. 1150–1157.
- Lu, Yijuan and Hao Jiang (2013). "Human movement summarization and depiction from videos". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6.
- Marcard, Timo von et al. (2018). "Recovering accurate 3d human pose in the wild using imus and a moving camera". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 601–617.
- McGinley, Jennifer L. et al. (2009). "The reliability of three-dimensional kinematic gait measurements: A systematic review". In: *Gait Posture* 29.3, pp. 360–369. ISSN: 0966-6362. DOI: <https://doi.org/10.1016/j.gaitpost.2008.09.003>. URL: <http://www.sciencedirect.com/science/article/pii/S0966636208002646>.
- McSweeney, Simon C, Lloyd F Reed, and Scott C Wearing (2020). "Reliability and minimum detectable change of measures of gait in children during walking and running on an instrumented treadmill". In: *Gait & posture* 75, pp. 105–108.
- Meldrum, Dara et al. (2014). "Test-retest reliability of three dimensional gait analysis: Including a novel approach to visualising agreement of gait cycle waveforms with Bland and Altman plots". In: *Gait & posture* 39.1, pp. 265–271.
- Merriaux, Pierre et al. (2017). "A study of vicon system positioning performance". In: *Sensors* 17.7, p. 1591.
- Mindler, Gabriel T et al. (2020). "Disease-specific gait deviations in pediatric patients with X-linked hypophosphatemia". In: *Gait & Posture* 81, pp. 78–84.
- Mitton, David et al. (2006). "3D reconstruction of the pelvis from bi-planar radiography". In: *Computer methods in biomechanics and biomedical engineering* 9.1, pp. 1–5.
- Mohandas Nair, Preeti, T George Hornby, and Andrea Louis Behrman (2012). "Minimal detectable change for spatial and temporal measurements of gait after incomplete spinal cord injury". In: *Topics in spinal cord injury rehabilitation* 18.3, pp. 273–281.

- Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee (2018). "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map". In: *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pp. 5079–5088.
- Müller, Jürgen and Michael Arens (2010). "Human pose estimation with implicit shape models". In: *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pp. 9–14.
- Newell, Alejandro, Kaiyu Yang, and Jia Deng (2016). "Stacked Hourglass Networks for Human Pose Estimation". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 483–499. ISBN: 978-3-319-46484-8.
- Nie, Xuecheng et al. (2018). "Human pose estimation with parsing induced learner". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2100–2108.
- Ning, G., Z. Zhang, and Z. He (2018). "Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation". In: *IEEE Transactions on Multimedia* 20.5, pp. 1246–1259. ISSN: 1520-9210. DOI: [10.1109/TMM.2017.2762010](https://doi.org/10.1109/TMM.2017.2762010).
- Núñez, Juan Carlos et al. (2019). "Multiview 3D human pose estimation using improved least-squares and LSTM networks". In: *Neurocomputing* 323, pp. 335–343. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2018.10.009>. URL: <http://www.sciencedirect.com/science/article/pii/S0925231218311858>.
- Ohashi, Akira et al. (2016). "Fisheye stereo camera using equirectangular images". In: *2016 11th France-Japan & 9th Europe-Asia Congress on Mechatronics (MECATRONICS)/17th International Conference on Research and Education in Mechatronics (REM)*. IEEE, pp. 284–289.
- Oleinikov, Georgii et al. (2014). "Task-based control of articulated human pose detection for openv1". In: *IEEE Winter Conference on Applications of Computer Vision*. IEEE, pp. 682–689.
- Omran, Mohamed et al. (2018). "Neural body fitting: Unifying deep learning and model based human pose and shape estimation". In: *2018 international conference on 3D vision (3DV)*. IEEE, pp. 484–494.
- Optical Motion Capture Guide* (2011). URL: http://physbam.stanford.edu/cs448x/old/Optical_Motion_Capture_Guide.html.
- Optitrack (2020). URL: <https://optitrack.com/hardware/compare/?products=600&601&603&602&291>.
- O'Connor, Ciara M et al. (2007). "Automatic detection of gait events using kinematic data". In: *Gait & posture* 25.3, pp. 469–474.
- Paterson, Kade, Keith Hill, and Noel Lythgo (2011). "Stride dynamics, gait variability and prospective falls risk in active community dwelling older women". In: *Gait & posture* 33.2, pp. 251–255.
- Pavlakos, Georgios, Xiaowei Zhou, and Kostas Daniilidis (2018). "Ordinal depth supervision for 3d human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7307–7316.
- Pavlakos, Georgios et al. (2018). "Learning to estimate 3D human pose and shape from a single color image". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459–468.
- Pavlakos, Georgios et al. (2019). "Expressive body capture: 3d hands, face, and body from a single image". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985.
- Pavlo, Dario et al. (2019). "3D human pose estimation in video with temporal convolutions and semi-supervised training". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762.
- Perry, J. (1992). *GAIT ANALYSIS Normal and Pathological Function*. SLACK Incorporated. ISBN: 978-1-55642-192-1.

- Pfister, Alexandra et al. (2014). "Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis". In: *Journal of medical engineering & technology* 38.5, pp. 274–280.
- Pillet, Hélène et al. (2014). "A reference method for the evaluation of femoral head joint center location technique based on external markers". In: *Gait & posture* 39.1, pp. 655–658.
- Pishchulin, Leonid et al. (2016). "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Podsiadlo, Diane and Sandra Richardson (1991). "The timed "Up & Go": a test of basic functional mobility for frail elderly persons". In: *Journal of the American geriatrics Society* 39.2, pp. 142–148.
- Pons-Moll, Gerard, David J Fleet, and Bodo Rosenhahn (2014). "Posebits for monocular human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2344.
- Pose Detection comparison (2019). URL: <https://www.learnopencv.com/pose-detection-comparison-wrncchai-vs-openpose/>.
- Presedo, Ana (2018). "Swing Phase Problems in Cerebral Palsy". In: *Handbook of human motion*, pp. 738–753.
- Qiu, Haibo et al. (2019). "Cross view fusion for 3D human pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4342–4351.
- Rafi, Umer and Juergen Gall (2016). "An Efficient Convolutional Network for Human Pose Estimation." In:
- Rayat Imtiaz Hossain, Mir and James J Little (2018). "Exploiting temporal information for 3d human pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 68–84.
- Reinschmidt, C et al. (1997). "Tibiofemoral and tibiocalcaneal motion during walking: external vs. skeletal markers". In: *Gait & Posture* 6.2, pp. 98–109.
- Roetenberg, Daniel (2006). *Inertial and magnetic sensing of human motion*. These de doctorat.
- Saleh, Michael and George Murdoch (1985). "In defence of gait analysis. Observation and measurement in gait assessment". In: *The Journal of bone and joint surgery. British volume* 67.2, pp. 237–241.
- Sandau, Martin et al. (2014). "Markerless motion capture can provide reliable 3D gait kinematics in the sagittal and frontal plane". In: *Medical engineering & physics* 36.9, pp. 1168–1175.
- Sapp, Benjamin, Alexander Toshev, and Ben Taskar (2010). "Cascaded models for articulated pose estimation". In: *European conference on computer vision*. Springer, pp. 406–420.
- Sarafianos, Nikolaos et al. (2016). "3d human pose estimation: A review of the literature and analysis of covariates". In: *Computer Vision and Image Understanding* 152, pp. 1–20.
- Sárándi, István et al. (2018). "Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation". In: *arXiv preprint arXiv:1809.04987*.
- Schepers, Martin, Matteo Giuberti, Giovanni Bellusci, et al. (2018). "Xsens mvn: Consistent tracking of human motion using inertial sensing". In: *Xsens Technologies*, pp. 1–8.
- Schmitz, Anne et al. (2014). "Accuracy and repeatability of joint angles measured using a single camera markerless motion capture system". In: *Journal of Biomechanics* 47.2, pp. 587–591. ISSN: 0021-9290. DOI: <https://doi.org/10.1016/j.jbiomech.2013.11.031>. URL: <http://www.sciencedirect.com/science/article/pii/S0021929013005848>.
- Schmitz, Anne et al. (2015). "The measurement of in vivo joint angles during a squat using a single camera markerless motion capture system as compared to a marker based system". In: *Gait Posture* 41.2, pp. 694–698. ISSN: 0966-6362. DOI: <https://doi.org/10.1016/j.gaitpost.2015.01.028>. URL: <http://www.sciencedirect.com/science/article/pii/S0966636215000314>.
- Schwartz, Michael H, Joyce P Trost, and Roy A Wurvey (2004). "Measurement and management of errors in quantitative gait data". In: *Gait & posture* 20.2, pp. 196–203.

- Sharma, Saurabh et al. (2019). "Monocular 3d human pose estimation by generation and ordinal ranking". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2325–2334.
- Shi, Yulong et al. (2018). "Fbi-pose: Towards bridging the gap between 2d images and 3d human poses using forward-or-backward information". In: *arXiv preprint arXiv:1806.09241*.
- Sholukha, Victor et al. (2006). "Double-step registration of in vivo stereophotogrammetry with both in vitro 6-DOFs electrogoniometry and CT medical imaging". In: *Journal of biomechanics* 39.11, pp. 2087–2095.
- Shrestha, Prarthana et al. (2009). "Synchronization of multiple camera videos using audio-visual features". In: *IEEE Transactions on Multimedia* 12.1, pp. 79–92.
- Sigal, Leonid, Alexandru O. Balan, and Michael J. Black (2009). "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion". In: *International Journal of Computer Vision* 87.1, p. 4. ISSN: 1573-1405. DOI: [10.1007/s11263-009-0273-6](https://doi.org/10.1007/s11263-009-0273-6). URL: <https://doi.org/10.1007/s11263-009-0273-6>.
- Sigal, Leonid, Alexandru O Balan, and Michael J Black (2010). "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion". In: *International journal of computer vision* 87.1-2, p. 4.
- Simon, Sheldon R. (2004). "Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems". In: *Journal of Biomechanics* 37 (12), pp. 1869–1880. DOI: <https://doi.org/10.1016/j.jbiomech.2004.02.047>. URL: <http://www.sciencedirect.com/science/article/pii/S0021929004001228>.
- Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.
- Springer, Shmuel and Galit Yogev Seligmann (2016). "Validity of the kinect for gait assessment: A focused review". In: *Sensors* 16.2, p. 194.
- Sun, Ke et al. (2019). "Deep high-resolution representation learning for human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703.
- Sun, Xiao et al. (2018). "Integral human pose regression". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 529–545.
- Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.
- Tajbakhsh, Nima et al. (2016). "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* 35.5, pp. 1299–1312.
- Tanaka, Ryo et al. (2018). "Validity of time series kinematical data as measured by a markerless motion capture system on a flatland for gait assessment". In: *Journal of Biomechanics* 71, pp. 281–285. ISSN: 0021-9290. DOI: <https://doi.org/10.1016/j.jbiomech.2018.01.035>. URL: <http://www.sciencedirect.com/science/article/pii/S002192901830071X>.
- Tang, Wei and Ying Wu (2019). "Does Learning Specific Features for Related Parts Help Human Pose Estimation?" In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1107–1116.
- Tang, Wei, Pei Yu, and Ying Wu (2018). "Deeply learned compositional models for human pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 190–206.
- Tang, Zhiqiang et al. (2018). "Quantized densely connected u-nets for efficient landmark localization". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 339–354.
- Tian, Yan et al. (2019). "Densely connected attentional pyramid residual network for human pose estimation". In: *Neurocomputing* 347, pp. 13–23. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2019.01.104>. URL: <http://www.sciencedirect.com/science/article/pii/S0925231219301973>.

- Toebes, Marcel JP et al. (2012). "Local dynamic stability and variability of gait are associated with fall history in elderly subjects". In: *Gait & posture* 36.3, pp. 527–531.
- Tome, Denis et al. (2018). "Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture". In: *2018 international conference on 3D vision (3DV)*. IEEE, pp. 474–483.
- Tu, Bo et al. (2013). "High precision two-step calibration method for the fish-eye camera". In: *Applied optics* 52.7, pp. C37–C42.
- Varol, Gul et al. (2017). "Learning from synthetic humans". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–117.
- Wang, Jue et al. (2019a). "Not All Parts Are Created Equal: 3D Pose Estimation by Modeling Bi-directional Dependencies of Body Parts". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7771–7780.
- Wang, Keze et al. (2019b). "3D human pose machines with self-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence*.
- Wang, Luyang et al. (2019c). "Generalizing monocular 3D human pose estimation in the wild". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0.
- Wang, Rui et al. (2019d). "Human pose estimation with deeply learned multi-scale compositional models". In: *IEEE Access* 7, pp. 71158–71166.
- Wang, Shijun and Ronald M Summers (2012). "Machine learning and radiology". In: *Medical image analysis* 16.5, pp. 933–951.
- Wei, Guoqiang et al. (2019). "View Invariant 3D Human Pose Estimation". In: *IEEE Transactions on Circuits and Systems for Video Technology*.
- Wei, Shih-En et al. (2016). "Convolutional Pose Machines". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Weygers, Ive et al. (2020). "Inertial sensor-based lower limb joint kinematics: A methodological systematic review". In: *Sensors* 20.3, p. 673.
- WHO (2008). *WHO (World Health Organization) global report on falls prevention in older age*. World Health Organization.
- Williamson, Sam et al. (2017). "Costs of fragility hip fractures globally: a systematic review and meta-regression analysis". In: *Osteoporosis International* 28.10, pp. 2791–2800.
- Wittwer, Joanne E et al. (2008). "Test–retest reliability of spatial and temporal gait parameters of people with Alzheimer's disease". In: *Gait & posture* 28.3, pp. 392–396.
- Wochatz, Monique et al. (2019). "Reliability and validity of the Kinect V2 for the assessment of lower extremity rehabilitation exercises". In: *Gait & Posture* 70, pp. 330–335.
- Wren, Tishya AL, Li-Shan Chou, and Thomas Dreher (2020). *Gait and Posture Virtual Special Issue "Clinical Impact of Instrumented Motion Analysis"*.
- Wren, Tishya AL et al. (2005). "Reliability and validity of visual assessments of gait using a modified physician rating scale for crouch and foot contact". In: *Journal of Pediatric Orthopaedics* 25.5, pp. 646–650.
- Wren, Tishya AL et al. (2011). "Efficacy of clinical gait analysis: A systematic review". In: *Gait & posture* 34.2, pp. 149–153.
- Wren, Tishya AL et al. (2013). "Outcomes of lower extremity orthopedic surgery in ambulatory children with cerebral palsy with and without gait analysis: results of a randomized controlled trial". In: *Gait & posture* 38.2, pp. 236–241.
- Wren, Tishya AL et al. (2020). "Clinical efficacy of instrumented gait analysis: Systematic review 2020 update". In: *Gait & Posture*.
- Wu, Ge et al. (2002). "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine". In: *Journal of biomechanics* 35.4, pp. 543–548.
- Xu, Xu et al. (2015). "Accuracy of the Microsoft Kinect™ for measuring gait parameters during treadmill walking". In: *Gait & posture* 42.2, pp. 145–151.

- Yang, Feng (2018). "Slip and Fall Risk Assessment". In: *Handbook of human motion*, pp. 656–677.
- Yang, Wei et al. (2017). "Learning feature pyramids for human pose estimation". In: *The IEEE International Conference on Computer Vision (ICCV)*. Vol. 2. 7.
- Yang, Yi and Deva Ramanan (2011). "Articulated pose estimation with flexible mixtures-of-parts". In: *CVPR 2011*. IEEE, pp. 1385–1392.
- Zeni Jr, JA, JG Richards, and JS Higginson (2008). "Two simple methods for determining gait events during treadmill and overground walking using kinematic data". In: *Gait & posture* 27.4, pp. 710–714.
- Zhang, Feng, Xiatian Zhu, and Mao Ye (2019a). "Efficient Human Pose Estimation in Hierarchical Context". In: *IEEE Access* 7, pp. 29365–29373.
- (2019b). "Fast human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3517–3526.
- Zhang, Hong et al. (2019a). "Human pose estimation with spatial contextual information". In: *arXiv preprint arXiv:1901.01760*.
- Zhang, Xiaoyan et al. (2019b). "3D human pose estimation via human structure-aware fully connected network". In: *Pattern Recognition Letters* 125, pp. 404–410. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2019.05.020>. URL: <http://www.sciencedirect.com/science/article/pii/S016786551830432X>.
- Zhao, Long et al. (2019). "Semantic graph convolutional networks for 3D human pose regression". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3425–3435.
- Zhou, Xingyi et al. (2017). "Towards 3d human pose estimation in the wild: a weakly-supervised approach". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 398–407.
- Zhu, Aichun et al. (2019). "Exploring hard joints mining via hourglass-based generative adversarial network for human pose estimation". In: *AIP Advances* 9.3, p. 035321.

Contribution à la capture du mouvement humain par stéréovision et machine learning pour l'analyse de la marche

Résumé

L'analyse de la marche est la mesure et l'évaluation de la capacité de marche qui peut être utilisée pour l'évaluation des risques de chute ou comme outil de diagnostic et de pronostic pour des applications cliniques. Toutefois, malgré la valeur clinique, plusieurs difficultés attribuées à l'instrumentation de référence actuelle, les systèmes de capture du mouvement basés sur des marqueurs, limitent l'utilisation à grande échelle dans les applications cliniques. Les systèmes actuels sont coûteux et nécessitent un environnement de laboratoire contrôlé. La procédure de test est également longue. L'élimination des marqueurs réduirait considérablement le temps de préparation du patient et serait plus efficace. L'objectif de cette étude est de concevoir un système de capture de mouvement sans marqueur pour les applications cliniques. Les appréciables progrès réalisés dans le domaine de la vision par ordinateur et en particulier dans celui des réseaux neuronaux convolutifs, ont permis de poursuivre cet objectif. Le système conçu se compose de quatre caméras RGB et peut estimer la position des centres communs grâce à une approche d'apprentissage profond. À cette fin, un nouvel ensemble de données spécifiques a été collecté, incluant des sujets asymptomatiques et pathologiques. Pour évaluer la validité du système développé, ses performances sont évaluées par rapport à un système de capture de mouvement basé sur des marqueurs en termes d'erreurs de position des articulations et de paramètres de marche cliniquement pertinents. Les résultats démontrent le potentiel élevé du système conçu pour des applications cliniques.

Mots clés – sans marqueur, analyse de la marche, estimation de la pose, apprentissage profond, machine learning, stéréovision.

Abstract

Gait analysis is the measurement of and ability evaluation of walking that can be used for fall risk assessment or as a diagnostic and prognostic tool for clinical applications. However, despite the clinical value, several difficulties attributed to the current established gold standard instrumentation, marker-based motion capture systems, limit the large-scale use in clinical applications. The current marker-based systems are costly and require a controlled laboratory environment. The test procedure is also time-consuming. Eliminating the markers would drastically shorten the patient preparatory time and would be more efficient. The objective of this study is to design a marker-less motion capture system for clinical applications. Recent advancements in computer vision and especially in convolutional neural networks, have provided the potential to pursue this objective. The designed system consists of four RGB cameras and can estimate the position of joint centers through a deep learning approach. For that purpose, a novel specific dataset has been collected including asymptomatic and pathologic subjects. To evaluate the validity of the developed system, its performance is assessed against a marker-based motion capture system in terms of joint position errors and clinically relevant gait parameters. The results demonstrate the high potential of the designed system for clinical applications.

Keywords – marker-less, gait analysis, pose estimation, deep learning, machine learning, stereovision.