Université de Montréal

Extraction de comportements reproductibles en avatar virtuel

Par

Kodjine Dare

Unité académique d'Informatique et de Recherche Opérationnelle, Faculté des Arts et Sciences

Mémoire présenté en vue de l'obtention du grade de Maîtrise

en Informatique, option Intelligence Artificielle

Octobre 2021

Université de Montréal

Département d'Informatique et de Recherche Opérationnelle, Faculté des Arts et Sciences

---

Ce mémoire intitulé

Extraction de comportements reproductibles en avatar virtuel

Présenté par

Kodjine Dare

A été évalué par un jury composé des personnes suivantes

Fabian Bastin

Président-rapporteur

Claude Frasson

Directeur de recherche

Esma Aimeur

Membre du jury

Résumé

Face à une image représentant une personne, nous (les êtres humains) pouvons visualiser les différentes parties de la personne en trois dimensions (tridimensionnellement – 3D) malgré l'aspect bidimensionnel (2D) de l'image. Cette compétence est maîtrisée grâce à des années d'analyse des humains. Bien que cette estimation soit facilement réalisable par les êtres humains, elle peut être difficile pour les machines. Dans ce mémoire, nous décrivons une approche qui vise à estimer des poses à partir de vidéos dans le but de reproduire les mouvements observés par un avatar virtuel. Nous poursuivons en particulier deux objectifs dans notre travail. Tout d'abord, nous souhaitons extraire les coordonnées d'un individu dans une vidéo à l'aide de méthodes 2D puis 3D. Dans le second objectif, nous explorons la reconstruction d'un avatar virtuel en utilisant les coordonnées 3D de façon à transférer les mouvements humains vers l'avatar. Notre approche qui consiste à compléter l'estimation des coordonnées 3D par des coordonnes 2D permettent d'obtenir de meilleurs résultats que les méthodes existantes. Finalement nous appliquons un transfert des positions par image sur le squelette d'un avatar virtuel afin de reproduire les mouvements extraits de la vidéo.

Mots-clés : Extraction de comportements, avatar virtuel, vidéo, simulation d'émotions, estimation pose 3D, reproduction de comportements.

Abstract

Given an image depicting a person, we (human beings) can visualize the different parts of the person in three dimensions despite the two-dimensional aspect of the image. This perceptual skill is mastered through years of analyzing humans. While this estimation is easily achievable for human beings, it can be challenging for machines. 3D human pose estimation uses a 3D skeleton to represent the human body posture. In this thesis, we describe an approach that aims at estimating poses from video with the objective of reproducing the observed movements by a virtual avatar. We aim two main objectives in our work. First, we achieve the extraction of initial body parts coordinates in 2D using a method that predicts joint locations by part affinities (PAF). Then, we estimate 3D body parts coordinates based on a human full 3D mesh reconstruction approach supplemented by the previously estimated 2D coordinates. Secondly, we explore the reconstruction of a virtual avatar using the extracted 3D coordinates with the prospect to transfer human movements towards the animated avatar. This would allow to extract the behavioral dynamics of a human. Our approach consists of multiple subsequent stages that show better results in the estimation and extraction than similar solutions due to this supplement of 2D coordinates. With the final extracted coordinates, we apply a transfer of the positions (per frame) to the skeleton of a virtual avatar in order to reproduce the movements extracted from the video.

Keywords: behavior extraction, virtual avatar, video, emotion simulation, 3D pose estimation, behavior reproduction.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

2D: Two-Dimensional

3D: Three-Dimensional

CNN: Convolutional Neural Networks

AP: Average Precision

SHG: Stacked Hourglass Networks

R-CNN: Regional CNN

COCO: Common Objects in Context

MPJPE: Mean Per Joint Position Error

LSTM: Long Short-Term Memory

JSON: JavaScript Object Notation

CSV: Comma-Separated Values

VGG: Visual Geometry Group

ResNet: Residual Network

RMPE: Regional Multi-person Pose Estimation

PAF: Part Affinity Fields

ReLU : Rectified Linear Unit

Remerciements

Je tiens tout d'abord à remercier mon directeur de recherche, Professeur Claude Frasson qui a toujours su se rendre disponible pour moi chaque fois que je rencontrais un problème ou que j'avais une question sur mes recherches. Il a permis que ce projet soit mon propre travail tout en m'orientant dans la bonne direction chaque fois qu'il pensait que j'en avais besoin.

Je suis reconnaissante envers Docteur Hamdi Ben Abdessalam pour m'avoir fourni assistance et suggestions tout au long de mes travaux. Sans son attention au détail, je n'aurais jamais pu élever mon travail à ce rang.

Je remercie le CRSNG-CRD (National Science and Engineering Research Council Cooperative Research Development), Prompt et BMU (Beam Me Up) pour le financement de ce travail.

Enfin, je tiens à remercier ma famille : mon père qui m'a transmis sa ténacité et le sens du travail bien fait, ma mère qui a toujours su comment me motiver, et mes précieuses sœurs qui n'ont jamais cessé de me soutenir.

Chapitre 1 – Introduction

1.1. Contexte général

Les êtres humains sont des créatures fortement dépendantes de leurs sens et parmi ces sens, ils sont largement dépendants de celui de la vue. En effet, l'interprétation des gestes et de la posture joue un rôle central dans l'apprentissage des émotions, mais aussi dans la socialisation. Étant donné une vidéo, le cerveau humain a la capacité de comprendre ce que les yeux voient sans aucune forme de conscience concernant le processus de reconnaissance. C'est ainsi que, confronté à une vidéo mettant en scène un individu effectuant des tâches, l'être humain a la capacité d'identifier simplement les différentes parties du corps de l'individu et inférer les parties non visibles. Si la réalisation de ce processus presque inconscient pour les êtres humains relève d'un apprentissage progressif avec l'âge, il a introduit une interrogation sur la capacité pour les ordinateurs de détecter automatiquement la pose du corps humain à partir d'images et de vidéos.

Avec les différentes percées technologiques de l'intelligence artificielle et l'augmentation des capacités computationnelles des ordinateurs, les limites du possible technologique ne cessent d'être repoussées. L'estimation de la pose humaine (HPE) est une branche de l'intelligence artificielle (plus précisément du domaine de la vision par ordinateur) qui se concentre sur la détection et l'estimation du corps humain sur des images et vidéos. À travers cette estimation de la pose humaine, l'objectif est de retrouver la position des points clés du corps humain (majoritairement les articulations) à travers l'analyse des médias (images et vidéos) afin d'en déduire le squelette et/ou la structure de l'humain ou des humains présents sur le média.

Cette tâche ayant pour but de fournir des informations sur la posture corporelle, les mouvements corporels et les gestes humains, constitue un véritable défi pour le domaine de la vision par ordinateur.

1.2. Objectifs

Nous avons articulé notre travail autour du transfert d'information basé sur l'estimation de la pose. C'est dans cette optique que nous avons dégagé deux objectifs pour notre travail. En

premier, nous nous concentrons sur l'estimation de la pose à partir de vidéos afin d'extraire des informations telles que les coordonnées des points clés du corps humain. L'extraction et l'entreposage de ces coordonnées fournissent une collection de données qui seront pertinentes pour la prochaine étape. Puis nous abordons le deuxième objectif de notre travail qui consiste à utiliser les coordonnées des points clés précédemment extraites à travers la réalisation du premier objectif pour la reproduction du comportement au moyen d'un avatar animé. Nous définissons l'animation d'un avatar comme une séquence de poses extraites dans une vidéo. Cet objectif vise à permettre de transférer les mouvements humains observés dans une vidéo vers un avatar animé afin de mettre en évidence la dynamique comportementale de l'humain à partir des points clés extraits.

La réalisation de ces deux objectifs peut conduire à de multiples implications. À travers notre travail, nous explorons une possible application de notre approche dans le cadre de la maladie d'Alzheimer. La maladie d'Alzheimer est un trouble cérébral irréversible et progressif qui détruit lentement la mémoire et les capacités de réflexion et affecte le comportement. Les patients atteints de la maladie d'Alzheimer peuvent avoir des comportements parfois spécifiques (marche, équilibre ou autre) qui pourraient être observés par caméra vidéo à différents moments de la journée, puis extraits et reconstruits dans un avatar virtuel. Cet avatar servirait de modèle de formation pour éduquer le personnel médical à reconnaître un épisode de patients atteints de la maladie d'Alzheimer et améliorer l'interaction avec ces derniers.

1.3. Organisation du mémoire

Dans la suite de ce mémoire, nous présentons au chapitre 2 les travaux pertinents dans le domaine de l'estimation de la pose humaine. Nous discutons des méthodes testées et des avantages de ces méthodes au chapitre 3. Le chapitre 4 est consacré à la méthodologie d'estimation et le chapitre 5 est destiné à la présentation des expérimentations et des résultats. Nous explorons la reconstruction des comportements en environnement virtuel au chapitre 6. Enfin, une conclusion générale termine notre mémoire et quelques perspectives sont proposées.

Chapitre 2 – État de l'art

Nous nous focalisons, dans cet état de l'art, sur les méthodes actuelles d'estimation de la pose humaine. Nous abordons aussi bien les méthodes d'estimation 2D que les méthodes 3D. Nous exposons les techniques utilisées par ces méthodes et leurs avancées.

Nous présentons des méthodes en faisant une claire distinction entre l'estimation bidimensionnelle (2D) et l'estimation tridimensionnelle (3D). Il existe deux catégories d'estimation de la pose 2D. Compte tenu du nombre de personnes à estimer sur une image, l'estimation de la pose peut être de la catégorie *estimation de personne unique* ou *estimation de personnes multiples*. Premièrement, nous présentons l'estimation de la pose 2D en faisons une distinction entre les méthodes de chaque catégorie. Ensuite, nous abordons l'estimation de la pose humaine 3D à travers les capteurs puis à travers les images et vidéos.

2.1 Réseau de Neurones à Convolution

Le réseau de neurones le plus utilisé dans le domaine de la vision par ordinateur pour la reconnaissance et classification d'images est le réseau de neurones à convolution dont la structure typique est illustrée à la figure 1. Dans certaines couches, le réseau applique des fonctions de convolution ou de mise en commun (pooling). Le réseau de neurone a convolution est une combinaison de couches convolutives et de pooling, de fonctions d'activation, telles que les unités linéaires rectifiées (ReLU) et les couches entièrement connectées.

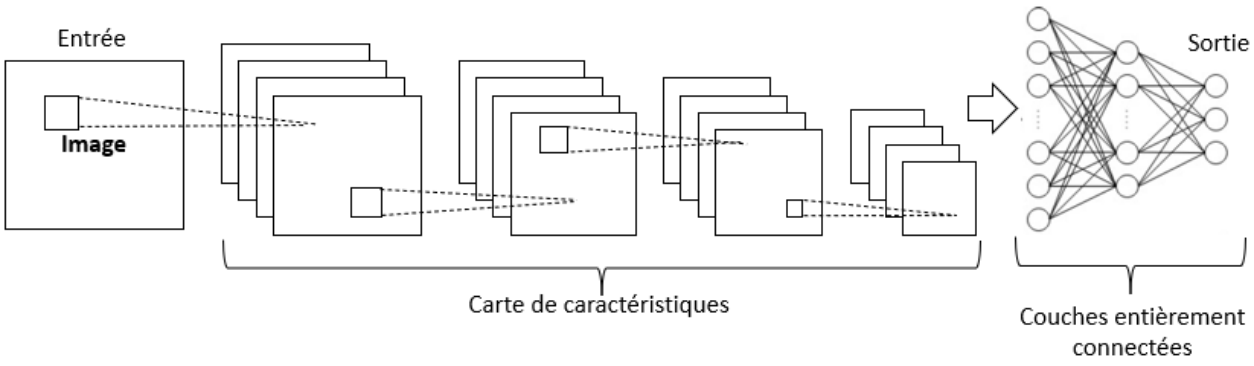

Figure 1. – Structure typique des réseaux de neurones à convolution

2.2 Estimation de la pose humaine en 2D

2.2.1 Personne unique

Les méthodes de cette catégorie sont destinées à estimer la pose humaine d'une seule personne sur une image. Selon la formulation de l'approche d'estimation de la pose, les méthodes peuvent être classées en deux sous-catégories : méthodes basées sur la régression et méthodes basées sur la détection.

Estimation par la régression

Ces méthodes réalisent l'estimation en faisant une régression directe des points clés du corps humain à partir des caractéristiques de l'image. Il existe plusieurs travaux qui valorisent cette approche.

Se reposant sur les fondements d'AlexNet [1] , Toshev et Szegedy [2] ont proposé l'une des premières méthodes d'estimation basée sur la régression. Leur méthode prônait un régresseur de réseau de neurones à convolution (CNN) nommé DeepPose.

L'introduction de cette approche et la qualité des résultats obtenus ont posé une base pour les méthodes basées majoritairement sur les réseaux de neurones à convolution plutôt que les approches classiques à apprentissage profond. Cependant, il est difficile d'apprendre les points clés directement à partir des caractéristiques sans autres procédures. Pour remédier à cette limitation, Carreira et al. [3] ont proposé un réseau avec retour d'erreurs itératives basé sur GoogleNet [4] qui traite récursivement la combinaison de l'image d'entrée et les résultats de sortie.

Sun et al. [5] ont introduit une méthode de régression sensible à la structure basée sur ResNet-50 [6]. Cette méthode utilise des articulations pour représenter la pose; une représentation basée sur l'ossature est conçue en tenant compte des informations sur la structure du corps humain pour obtenir des résultats plus stables que d'utiliser uniquement des positions articulaires.

Un modèle gérant plusieurs tâches étroitement liées au corps humain (par exemple, reconnaissance de marche couplée à la reconnaissance de gestes de patients Alzheimer) peut

apprendre diverses caractéristiques (ou traits) d'une image pour améliorer la prédiction des coordonnées des points clés. En effet, en partageant des représentations entre des tâches connexes comme estimation de pose et reconnaissance d'actions basée sur la pose, le modèle apprend les représentations pertinentes sur la localisation de certaines parties du corps lors de la reconnaissance d'actions, fournissant ainsi un cadre de régulation pour mieux réaliser la tâche relative à l'estimation de la pose. C'est dans cette optique que Li et al. [7] ont utilisé un cadre multitâche de type AlexNet pour gérer la tâche de prédiction de coordonnées des parties du corps avec des images complètes, par la régression, et la tâche de détection de partie du corps à partir de portions (patchs) d'images obtenues par la technique de fenêtre glissante. Par la suite, Gkioxari et al. [8] ont utilisé une architecture R-CNN pour détecter une personne sur une image, estimer la pose de cette personne et classer l'action réalisée par la personne.

Fan et al. [9] ont proposé un modèle de CNN profond à doubles sources (c.-à-d. portions d'image et images complètes) en entrée, pour déterminer si une portion d'image contient une partie du corps et trouver son emplacement exact dans la portion. La combinaison de deux tâches conduit à des résultats améliorés.

Estimation par la détection

Les méthodes basées sur la détection visent à entraîner un détecteur des parties du corps humain afin de prédire les coordonnées des points clés du corps.

Certains travaux abordent l'estimation de la pose sous forme de prédiction de cartes thermiques (heatmaps). Wei et al. [10] ont proposé une solution basée sur les réseaux à convolution nommée Convolutional Pose Machines pour régresser les cartes thermiques en plusieurs étapes et en utilisant une supervision intermédiaire pour éviter la disparition du gradient.

Plutard, Newell et al. [11] ont introduit un réseau encodeur-décodeur de sabliers empilés (staked hourglass - SHG) qui se compose d'étapes consécutives "pooling" (réduction de la taille d'image en préservant les caractéristiques) et de suréchantillonnage des couches pour capturer des informations à toutes les échelles. Depuis lors, des variations complexes de cette architecture ont été développées pour estimation de la pose (Chu et al. [12], Yang et al. [13]).

Les travaux de Chou et al. [14] ont mis en lumière un réseau basé sur l'apprentissage avec deux mêmes réseaux de sabliers empilés comme générateur et discriminateur respectivement. Le générateur prédit l'emplacement de la carte thermique de chaque point clé, tandis que le discriminateur distingue les cartes thermiques de données de terrain des cartes thermiques générées.

Tang et al. [15] ont construit un réseau de supervision basé sur un réseau de sabliers, pour décrire les relations complexes et réalistes entre les parties du corps et apprendre les informations sur le modèle de composition (l'orientation, l'échelle et la forme de chaque partie du corps) dans les corps humains.

2.2.2 Personnes multiples

Les méthodes de cette catégorie sont destinées à estimer la pose humaine de plusieurs personnes sur une image. L'estimation de personnes multiples diffère de celle précédemment présentée du fait de sa complexité. En effet, compte tenu de la multiplicité des poses à estimer, les différentes méthodes doivent gérer à la fois les tâches de détection et de localisation, car il n'y a pas d'indication du nombre de personnes dans les images d'entrée. Selon l'approche adoptée par les différentes méthodes, elles peuvent être classées comme descendantes (top-down) et méthodes ascendantes (bottom-up).

Méthodes descendantes

Les méthodes descendantes abordent l'estimation de la pose par l'usage des détecteurs de personne pour obtenir un ensemble de boîtes de délimitations (bounding box) contenant une personne différente dans l'image d'entrée. Puis, elles exploitent directement les estimateurs de pose de personne unique existant afin de prédire les poses humaines.

La majorité des travaux s'est concentrée sur l'estimation des parties du corps basé sur des détecteurs humains existants tels que Faster R-CNN (Ren et al. [16]), Mask R-CNN (He et al.[17]).

L'estimation des poses en présence d'occlusion et de troncature se produit souvent dans des images contenant plusieurs personnes, car le chevauchement des membres est inévitable. Les détecteurs humains peuvent échouer lors de la première étape des méthodes descendantes

(détection de boîtes de délimitations contenant une personne différente) en raison d'une occlusion ou d'une troncature. Par conséquent, la robustesse à l'occlusion ou à la troncature est un aspect important de ce type l'estimation. De ce fait, Iqbal et Gall [18] ont utilisé un estimateur de pose basé sur des machines à convolutions [10] pour générer des poses initiales. Ils appliquent ensuite, la programmation linéaire en nombre entiers (Integer Linear Programming) pour obtenir les poses finales.

Par la suite, Fang et al. [19] ont adopté le réseau de transformateurs spatiaux (Spatial Transformer Network), la suppression sans maximum (Non-Maximum Suppression) et un réseau de sabliers pour faciliter l'estimation de la pose en recalculant une région (de grande précision) pour chaque personne dans le cas où les boîtes de délimitations sont inexactes. Xiao et al. [20] ont ajouté plusieurs couches de déconvolutions sur la dernière couche de convolution de ResNet pour générer des cartes thermiques à partir de caractéristiques de haute et basse résolution. Chen et al. [21] ont proposé un réseau de pyramides en cascade en utilisant des cartes de caractéristiques multi-échelles de différentes couches pour obtenir plus d'inférence à partir de caractéristiques locales et globales. Sur la base de distributions d'erreurs de pose similaires de différentes approches d'estimation, Moon et al. [22] ont conçu le filet (net) "PoseFix" pour affiner les poses estimées à partir de toutes méthodes.

Les performances de ce type de méthodes sont affectées par les résultats de la détection des personnes.

Méthodes ascendantes

Les méthodes ascendantes prédisent directement tous les points clés de toutes les personnes présentes sur l'image, puis les regroupent en squelettes distincts. Avec les méthodes descendantes, le nombre de personnes dans l'image d'entrée affecte directement le temps de calcul.

Les méthodes ascendantes se composent de deux étapes principales : l'extraction des caractéristiques locales pour la détection des parties du corps des individus (1) et l'assemblage des parties détectées par individus (2).

Pishchulin et al. [23] ont proposé un détecteur de parties du corps basé sur Fast R-CNN appelé DeepCut, qui est l'une des premières approches ascendantes en deux étapes. Il détecte d'abord toutes les parties du corps des individus, puis étiquette chaque partie et assemble ces parties en utilisant la programmation linéaire entière pour une pose finale.

Insafutdinov et al. [24] ont proposé DeeperCut en appliquant un détecteur de partie du corps plus fort avec une meilleure stratégie d'optimisation incrémentale conduisant à des performances améliorées ainsi qu'à une vitesse plus rapide. Cao et al. [25] ont proposé une méthode efficace qui utilise une représentation non paramétrique appelée « Part Association Fields » (PAF), un ensemble de vecteurs 2D qui codent l'emplacement et l'orientation des membres sur le domaine de l'image.

Plus tard, ils ont étendu leurs travaux à la construction d'un détecteur nommé OpenPose [26] en simplifiant le nombre de convolutions. Motivé par OpenPose et la structure de sablier empilés, Newell et al. [27] a introduit un réseau profond à un étage pour obtenir simultanément des détections et groupements. Suite à cette approche, Jin et al. [28] ont proposé une nouvelle méthode de groupement hiérarchique de graphes pour apprendre le groupement des parties humaines.

Certaines méthodes ont utilisé des structures multitâches. C’est le cas de celle suggérée par Papandreou et al. [29] appelé PersonLab pour combiner le module d'estimation de pose et le module de segmentation de personne pour la détection et l'association de points clés. PersonLab prédit de manière synchrone les cartes thermiques conjointes de tous les points clés pour chaque personne et leurs déplacements relatifs. Ensuite, le regroupement commence à partir de la détection la plus sûre avec un processus de décodage gourmand basé sur un graphe basé sur un squelette humain.

Les méthodes ascendantes sont généralement plus rapides que les méthodes descendantes, car elles détectent directement tous les points clés et les regroupent en poses individuelles à l'aide de stratégies d'association de points clés.

2.3 Estimation de la pose humaine en 3D

L'estimation de la pose humaine en 3D présente beaucoup plus de défis que l'estimation 2D ~~due~~ en raison de la problématique du troisième axe. En effet, l'estimation 3D nécessite les coordonnées de l'axe z, coordonnées qui ne sont pas disponibles dans les images bidimensionnelles. Cette contrainte liée à l'axe Z a engendré l'approche de l'estimation 3D sous différents angles, créant ainsi plusieurs catégories pour les méthodes.

Certaines méthodes utilisent des sources de données comme : capteurs de profondeur [30, 31, 32], capteurs de nuages de points [33, 34, 35], unités de mesure d'inertie [36, 37, 38], appareil à radiofréquence [39] et autres capteurs [40, 41, 42]. D'autres exploitent des images avec multiples champs de vision [43, 44, 45, 46].

Bien que ces méthodes obtiennent des résultats satisfaisants, toutes ces approches fonctionnent dans des environnements bien spécifiques ou nécessitent des marqueurs spéciaux sur le corps humain. Les objectifs de nos travaux sont premièrement d'extraire les comportements observés dans une vidéo (capturée par une seule caméra sans certaines conditions d'enregistrement ou équipements particuliers) et de reproduire ensuite ces comportements. Ainsi, les méthodes citées plus haut ne sont pas appropriées. De ce fait, nous parlerons plus des méthodes axées sur les images à champs de vision unique provenant de caméras monoculaires.

2.3.1 Approches basées sur les caméras monoculaires

Dans l'estimation 3D de la pose d'une seule personne, certaines méthodes adoptent **des modèles de corps humain** (modèle basé sur le squelette, modèle plan ou modèle volumétrique) pour déduire et reconstruire un humain en 3D tandis que d'autres n'y ont pas recours. Sur la Figure 2 nous montrons à quoi ressemble ces différents modèles de corps humains.

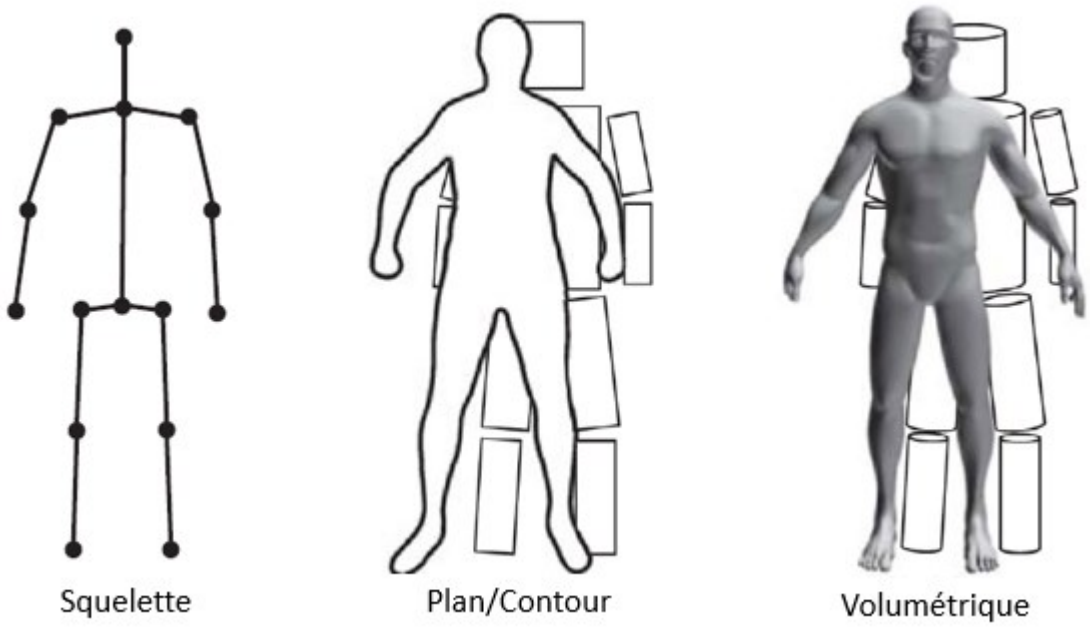

Figure 2. – Modèles humains communément utilisés dans la modélisation

Parmi les méthodes qui ne reposent pas sur les modèles de corps humain, plusieurs travaux ont proposé d'estimer la pose 3D directement à partir des caractéristiques des images soumises. Li et al. [47] ont proposé une approche où les paires image-pose 3D ont été utilisées en entrée de réseau. Un réseau de scores peut attribuer des scores élevés aux paires de poses image-3D correctes et des scores faibles aux autres paires. Pavlakos et al. [48] ont formé le réseau avec des profondeurs ordinales supplémentaires d'articulations humaines en tant que contraintes, par lesquelles les jeux de données humains 2D peuvent également être ajoutés avec des annotations de profondeurs ordinales.

À part ces méthodes qui réalisent une estimation 3D directement à partir des images 2D, il existe aussi des méthodes qui infèrent la pose 3D paramétrée sur la projection de coordonnées 2D préalablement estimées. C'est le cas, par exemple de Martinez et al. [49] qui ont proposé une solution simple avec réseau résiduel entièrement connecté efficace pour régresser les emplacements d'articulation 3D en fonction des emplacements d'articulation 2D. Similairement, Tekin et al. [50] et Zhou et al. [51] ont utilisé des cartes thermiques 2D au lieu de la pose 2D comme représentations intermédiaires pour estimer la pose 3D. Jahangiri et Yuille [52], Sharma

et al. [53], et Li et Lee [54] ont d'abord généré plusieurs hypothèses de pose 3D diverses puis appliqué des réseaux de classement pour sélectionner la meilleure pose 3D.

Contrairement aux méthodes précédentes, les méthodes basées sur des modèles utilisent un modèle de corps humain paramétrique pour estimer la pose et la forme humaines à partir d'images. Certains travaux ont utilisé le modèle corporel de Loper et al. [55] nommé SMPL: A Skinned Multi-Person Linear Model et ont tenté d'estimer les paramètres 3D à partir d'images. Par exemple, Bogo et al. [56] ont ajusté le modèle SMPL aux coordonnées 2D estimées et ont proposé une méthode basée sur l'optimisation pour récupérer les paramètres SMPL à partir des joints 2D. Tan et al. [57] ont déduit des paramètres SMPL en entraînant d'abord un décodeur pour prédire des silhouettes à partir de paramètres SMPL avec des données synthétiques, puis en apprenant un encodeur d'image avec le décodeur entraîné. L'encodeur formé peut prédire les paramètres SMPL à partir des images d'entrée. L'apprentissage direct des paramètres de SMPL est difficile, certains travaux ont employé des intermédiaires comme la pose 2D intermédiaire et la segmentation du corps humain (Pavlakos et al. [58]), la segmentation des parties du corps (Omran et al. [59], Varol et al. [60]). Afin de surmonter le problème du manque de données d'entraînement pour le modèle du corps humain, Kanazawa et al. [61] ont utilisé un générateur pour prédire les paramètres de SMPL et un discriminateur pour distinguer le modèle SMPL réel et prédit. Mehta et al. [62] ont prédit les emplacements relatifs des articulations à partir de cartes thermiques 2D suivant le modèle basé sur le squelette de corps. Nie et al. [63] ont utilisé LSTM pour exploiter les localisations d'articulations 2D globales et les images de parties de corps locales en suivant un modèle basé sur le squelette de corps (deux indices pour l'estimation de la profondeur des articulations). Zhou et al. [64] ont intégré un modèle basé sur le squelette dans un réseau pour l'estimation générale de la pose d'objet articulé qui fournit des indices d'orientation et de rotation. Cheng et al. [65] ont introduit un modèle Cylinder Man pour générer des étiquettes d'occlusion pour les données 3D et effectué une augmentation des données. Un terme de régularisation de pose a été introduit pour pénaliser les mauvaises étiquettes d'occlusion estimées. Xiang et al. [66] ont utilisé le modèle d'Adam [67] pour reconstruire les mouvements 3D. Wang et al. [68] ont présenté un nouveau modèle au niveau du maillage des os

humains, qui découple la modélisation osseuse et les variations en définissant les longueurs osseuses et les angles articulaires.

Les méthodes axées sur l'estimation des images peuvent tout autant être utilisées pour l'estimation vidéo que pour une succession d'images (étant donné qu'une vidéo représente une succession d'images). Néanmoins, certaines méthodes sont plus propices que d'autres pour cette tâche. Parmi ces méthodes, la plupart des approches adoptent une démarche en deux étapes : obtenir d'abord une reconstruction 3D, puis traitement du résultat par lissage en résolvant un problème d'optimisation [69, 70, 71, 72, 73].

Cette étude de l'existant nous a permis de prendre connaissance de l'entendu des moyens disponibles pour estimer la pose humaine tant bidimensionnelle que tridimensionnelle. Nous avons pu relever que les méthodes d'estimation tridimensionnelle sont les plus pertinentes pour réaliser une reproduction de comportement optimale. En effet compte tenu de notre objectif de reproduction au moyen d'un avatar, les coordonnées nécessaires pour ce genre de tâche doivent être disponibles en 3D. Grâce à cette étude de l'art, nous avons trouvé les méthodes d'estimation 3D qui présentent les meilleurs résultats, mais aussi des ouvertures pour la réalisation de nos objectifs. Les approches permettant l'estimation 3D par la projection probabiliste des coordonnées 2D [49, 50, 51, 52] proposent de bons résultats tout en valorisant des résultats satisfaisants sur jeux de données évalués. Ces approches n'exploitent pas des modèles de corps humain (exemple modèle volumétrique). D'autre part, nous avons aussi observé que les méthodes s'appuyant sur les modèles de corps humain offraient également de bons résultats. Plus précisément, les méthodes reposant sur le modèle SMPL offrent divers avantages en plus des résultats satisfaisants sur les données arbitraires. En effet, certaines méthodes [56, 57, 58, 59, 61]) permettent d'obtenir non seulement les coordonnées des points clés humains suivants des modèles réalistes humains, mais aussi une estimation de la forme de l'individu.

2.4 Récapitulatif de l'état de l'art

Ces dernières années, avec l'apprentissage en profondeur montrant de bonnes performances sur de nombreuses tâches de vision par ordinateur telles que la classification d'images, la segmentation, la détection d'objets, etc., l'estimation de la pose humaine réalise également des

progrès rapides en utilisant la technologie d'apprentissage en profondeur. Les principaux développements incluent des réseaux bien conçus avec une grande capacité d'estimation, des ensembles de données plus riches et une exploration plus pratique des modèles corporels. Nous avons fait l'état de l'art des méthodes d'estimation de la pose humaine 2D et 3D basées sur l'apprentissage profond à partir d'images monoculaires ou de séquences vidéo d'humains. Nous avons aussi listé des approches d'estimations de la pose 3D basées sur des capteurs tels que des capteurs de profondeur, source de lumière infrarouge et la radiofréquence. Le tableau 1 résume les méthodes dont nous avons parlé avec leur classification.

Tableau 1. – Classification des méthodes

Estimation	**Catégories**	**Sous-catégories**	**Méthodes**
2D (Méthodes monoculaires)	**Personnes Uniques**	Estimation par régression	AlexNet [1] , Toshev et Szegedy [2], Carreira et al. [3], GoogleNet [4], Sun et al. [5], Li et al. [7], Gkioxari et al. [8], Fan et al. [9]
		Estimation par détection	Wei et al. [10], Newell et al. [11], Chu et al. [12], Yang et al. [13], Chou et al. [14], Tang et al. [15]
	Multi-personnes	Estimation Descendantes	Ren et al. [16],He et al.[17]), Iqbal et Gall [18], Fang et al. [19], Xiao et al. [20], Chen et al. [21], Moon et al. [22]
		Estimation Ascendantes	Pishchulin et al. [23], Insafutdinov et al. [24], Cao et al. [25], OpenPose [26] ,Newell et al. [27], Jin et al. [28], Papandreou et al. [29]
3D (Basée sur les capteurs)	**Capteurs de profondeur**		Yu et al. [30], Kadkhodamohammadi et al. [31], Zhi et al. [32]
	Capteurs de nuages de points		Charles et al. [33], et al. [34], Wang et al. [35]
	Unités de mesure d'inertie		Marcard et al. [36], Zhang et al. [37], Huang et al. [38]
	Autres capteurs		Isogawa et al.[40], Tome et al. [41], Saini et al. [42]

3D (Méthodes monoculaires)	**Indépendants des modèles humains**	Estimation direct	Li et al. [47], Pavlakos et al. [48]
		Inférence 2D en 3D	Martinez et al. [49], Tekin et al. [50], Zhou et al. [51], Jahangiri et Yuille [52], Sharma et al. [53],Li et Lee [54]
	Basés sur les modèles humains	Modèle volumétrique	Bogo et al. [56], Tan et al. [57], Pavlakos et al. [58], Omran et al. [59], Varol et al. [60], Kanazawa et al. [61], Cheng et al. [65] , Wang et al. [68]
		Modèle squelettique	Mehta et al. [62], Nie et al. [63], Zhou et al. [64],

L'un des objectifs de nos travaux étant la reconstruction de la pose humaine à l'aide d'un avatar, nous nous concentrons premièrement sur les méthodes d'estimation de la pose en 3D.

Parmi toutes les sous-catégories d'approches d'estimation de la pose 3D présentées plus haut, les sous-catégories les plus pertinentes pour nos travaux sont les méthodes d'estimation de la pose 3D par:

- Inférence 2D en 3D
- Modèle volumétrique

Pour commencer, nous allons évaluer une méthode de chacune de ces sous-catégories. Ainsi, pour la sous-catégorie « Inférence 2D en 3D » nous choisissons d'évaluer la méthode proposée par **Martinez et al. [49]**. Cette méthode constitue l'une des plus solides parmi les méthodes d'inférence 3D à partir de 2D. En ce qui concerne la sous-catégories « Modèle volumétrique », nous choisissons d'évaluer l'approche de **Kanazawa et al. [61]**.

Cette évaluation a pour but de trouver, non seulement, l'approche qui généralise le mieux sur les données arbitraires, mais aussi l'approche la plus applicable à nos travaux. Dans le chapitre suivant nous présentons aussi bien les jeux de données et les métriques utilisées durant notre estimation, que les méthodes elles-mêmes.

Chapitre 3 – Méthodes évaluées

Cette section décrit principalement les techniques qui ont été testées et les jeux de données utilisés afin de choisir la meilleure approche pour réaliser les travaux. La détermination de la meilleure méthode repose sur l'obtention d'erreur minimale sur le jeu d'évaluation et surtout l'habilité d'une méthode à généraliser sur des données arbitraires de différents formats. Un autre facteur que nous considérons dans le choix de la méthode est la robustesse face à une optimisation. Comme présenté dans la section précédente, l'estimation 3D et 2D présentent différentes approches pour réaliser les tâches respectives. Nous allons commencer par évaluer des méthodes d'estimation 3D. Nous présentons également les jeux de données sur lesquelles nous avons effectué l'évaluation quantitative et les métriques d'évaluation utilisées

3.1 Jeux de données et métrique d'évaluation

3.1.1 Jeux de données

Human3.6M

C'est le jeu de données [74] le plus largement utilisé pour l'estimation de la pose 3D à partir d'images et de vidéos monoculaires. Il dépeint 11 acteurs professionnels (6 hommes et 5 femmes) effectuant 17 activités (à savoir, fumer, prendre des photos, marcher) à partir de 4 vues différentes dans un environnement de laboratoire intérieur. Cet ensemble de données contient 3,6 millions de poses humaines 3D avec annotation de données de terrain 3D basées sur des marqueurs de haute précision. Il existe 3 protocoles avec différents fractionnements de données d'entraînement et de test. Le protocole n ° 1(P1) utilise des images des sujets S1, S5, S6 et S7 pour l'entraînement et des images des sujets S9 et S11 pour les tests. Le protocole n ° 2(P2) utilise le même partage de tests d'entraînement que P1, mais les prédictions sont ensuite traitées par une transformation rigide avant de les comparer aux données de terrain. Le protocole n ° 3(P3) utilise des images des sujets S1, S5, S6, S7 et S9 pour l'entraînement et des images des sujets S11 pour les tests.

Microsoft Common Objects in Context

Le jeu de données Microsoft Common Objects in Context (COCO) [75] constitue l'un des plus utilisés dans le contexte de l'estimation 2D. Il a été initialement proposé pour la détection et la segmentation d'objets dans les environnements naturels. Avec des améliorations et des extensions, l'utilisation de COCO couvre le sous-titrage d'images et la détection des points clés. Les images sont collectées à partir de la recherche d'images Google, Bing et Flickr avec des catégories d'objets isolées ou par paires. Il contient plus de 330000 images et 200000 sujets étiquetés avec des points clés, et chaque personne est étiquetée avec 17 articulations. Un fichier JSON est utilisé pour stocker ces annotations. L'ensemble de données COCO a apporté à la table un mélange très intéressant de données, avec diverses poses humaines utilisées dans différentes échelles corporelles, contenant également des modèles d'occlusion, avec des environnements sans contraintes. Cet ensemble de données est divisé en ensemble d'entraînement (train), de validation (val) et de test (test).

3.1.2 Métrique d'évaluation

Précision moyenne

La précision moyenne (average precision – AP) est un indice permettant de mesurer l'exactitude de la détection des points clés en fonction de la précision (le rapport des vrais résultats positifs au total des résultats positifs) et du rappel (le rapport des vrais résultats positifs au nombre total de vrais positifs au sol). Cette mesure calcule la valeur de précision moyenne pour le rappel sur 0 à 1. Il existe plusieurs variantes similaires, c'est le cas par exemple de la précision moyenne pour un K nombre de points clés estimés. Ce dernier est connu sous le nom de APK – Average Precision of Keypoints.

Mean Per Joint Position Error (MPJPE)

L'erreur de position moyenne par point clé (MPJPE) est la mesure la plus largement utilisée pour évaluer les performances de l'estimation de pose 3D. Elle calcule la distance euclidienne entre les points clés 3D estimés et la donnée de terrain en millimètres, moyennée sur toutes les articulations dans une seule image. Dans le cas d'une séquence d'images, l'erreur moyenne est moyennée sur toutes les images. Pour différents jeux de données et différents protocoles, il existe

différents traitements des données des points clés estimés avant de calculer le MPJPE. Pour N nombre de points clés, J_i la donnée de terrain et J_i^* la prédiction de la position du point i -ème point clé, cette métrique est calculée comme suit :

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \| J_i - J_i^* \|_2 \qquad (1)$$

3.2 Technique d'estimation de la pose humaine en 3D

Afin de déterminer l'approche la plus convenable pour extraire les coordonnées utiles pour la reproduction de comportement, nous avons évalué des méthodes existantes. Nous avons choisi des méthodes qui offrent des résultats satisfaisants et proposant deux approches différentes.

Nous avons décidé d'évaluer une méthode qui repose sur l'inférence de coordonnées 3D à partir de coordonnées 2D indépendantes d'un modèle de corps humains [49] et une méthode qui estime les coordonnées 3D par régression directe de l'image basée sur un modèle volumétrique de corps humain [61]. Ce choix de méthodes nous a permis d'évaluer non seulement les méthodes elles-mêmes et leurs performances, mais aussi les catégories de méthodes.

3.2.1 A Simple yet Effective Baseline for 3d Human Pose Estimation

Proposée par Martinez et al., cette méthode constitue une référence simple pour l'estimation 3D de la pose humaine. Le choix de cette méthode pour évaluation a été motivé par la simplicité que prône l'approche comparée à la qualité des résultats reportés par les travaux. Elle consiste à inférer directement les coordonnées 3D à partir de coordonnées 2D obtenues avec une architecture d'estimation de pose humaine 2D.

Nous avons mis sur pied une architecture conforme à la structure présentée par les auteurs de la méthode (Martinez et al. [49]): un réseau à convolution simple à deux couches avec normalisation par lots et une fonction d'activation de type ReLU (rectification linéaire) [76]. En entrée, le modèle est alimenté avec les prédictions 2D comme prescrit par Newell et al. [11].

Sur le jeu de donnée Human3.6M, nous avons observé une erreur de position moyenne par point clé (MPJPE) de 73,6 millimètres pour le protocole d'évaluation P1 (déjà cité au niveau des métriques d'évaluation). Toujours sur le même jeu de données, nous avons aussi relevé des

résultats qualitatifs à la mesure du MPJPE observé. Cependant, ces résultats aussi impressionnants sur ce jeu de donnée ont été contrastés par la performance de la méthode sur des images plus générales (n'appartenant pas au jeu de donnée Human3.6M).

Nous avons observé que l'estimation 3D était d'autant plus éloignée de la réalité sur les images à forte occlusion de pixel et sur lesquelles l'individu dont la pose est à estimer n'est pas centré, comme le montre la Figure 3. L'évaluation approfondie de l'approche nous a permis d'attribuer ces limitations à la forte dépendance du modèle à un estimateur de coordonnées 2D[48]. En effet, puisque l'estimation 3D repose entièrement sur une première estimation 2D et non sur l'image elle-même, une mauvaise estimation des coordonnées 2D entraîne automatiquement une mauvaise estimation des coordonnées 3D.

3.2.2 End-to-end Recovery of Human Shape and Pose

Cette approche permet de déduire un maillage 3D directement à partir d'une image, par opposition à l'approche précédente qui utilise la détermination des points clés 2D comme intermédiaire. Nous avons reproduit l'architecture présentée par Kanazawa et al. [61] afin d'évaluer les résultats. Le modèle établi comporte un *encodeur*, un *régresseur* et un *discriminateur*. Nous soumettons une image à un encodeur à convolution, puis les caractéristiques de l'image sont envoyées au régresseur dont l'objectif est de déduire la forme du corps humain, la pose 3D et les paramètres de la caméra. La pose et la forme humaine déduite après la régression 3D sont également envoyées à un réseau discriminateur dont la tâche est de déterminer si les paramètres 3D représentent la forme et la pose réaliste (et non une forme improbable) d'un être humain.

Nous avons observé sur le jeu de données Human3.6, une erreur de position moyenne par point clé de 87,9 millimètres suivant le protocole d'évaluation P1. Nous avons aussi mesuré les résultats qualitatifs du modèle. Sur des images du jeu de données COCO, le modèle a obtenu de meilleurs résultats que la méthode précédente et s'est avéré meilleur à la généralisation. De plus, nous avons constaté que cette approche évite la nécessité d'une estimation intermédiaire (inférence de coordonnées 3D à partir de coordonnée 2D) empêchant ainsi la perte des informations précieuses dans l'image. Similairement à la méthode évaluée précédemment, le modèle ne

fournissait pas de bons résultats sur les images non contrées sur un individu en particulier (Figure 3).

3.2.3 Observations faites

Nous constatons que la dernière approche évaluée, celle suggérée par Kanazawa et al. est plus robuste sur les images arbitraires que l'approche de Martinez et al. Cette méthode (Kanazawa et al.) se basant directement sur la régression de l'image, elle permet de conserver les informations clés sur les traits (ou caractéristiques) des images. Cela permet d'obtenir de meilleures estimations. L'approche de Marinez et al. fourni certes une erreur (de position moyenne par point clé) inférieure, cependant, elle repose entièrement sur une phase intermédiaire d'une estimation de point 2D. Cette dépendance à cette estimation 2D entraîne de mauvaises prédictions 3D dans certains cas. En effet, si les données 2D intermédiaires sont faussées, l'estimation 3D qui en résulte est automatiquement incorrecte.

L'approche de reconstruction de maillage semble meilleure que celle de Martinez et al. [49]pour plusieurs raisons :

- La méthode permet d'obtenir plus que de simples coordonnées 3D, mais aussi une reconstruction de forme
- La méthode évite une perte d'information en faisant la régression directement à partir de l'image

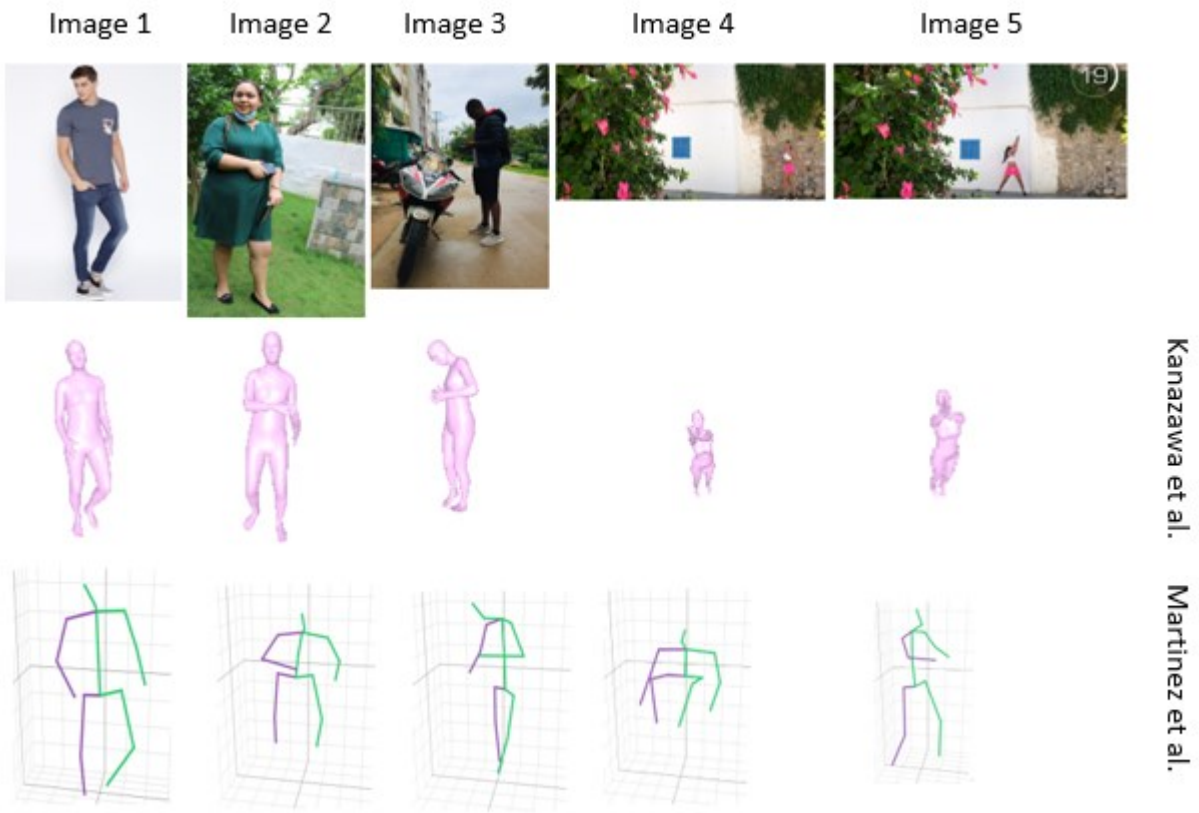

Figure 3. – Comparaison d'estimation 3D

Nous observons des limitations sur certaines images, plus précisément sur les images décentrées sur l'individu. Ces limitations s'expliquent par le fait que la méthode n'utilise pas des boîtes de délimitations. L'utilisation de boîtes de délimitations permet de mieux régresser la pose à partir des boîtes de délimitations.

Pour résoudre ces limitations nous décidons de rajouter des boîtes de délimitations à la méthode de maillage suggérer par Kanazawa et al. Pour ce faire, nous décidons d'utiliser ces boîtes de délimitations obtenues à partie de prédiction de la position d'un individu sur une image.

Pour cette prédiction de la position, nous allons considérer des méthodes d'estimation 2D dont les estimations serviront à calculer des boîtes de délimitations. Ainsi nous décidons d'évaluer un certain nombre de méthodes d'estimation de la pose 2D.

3.3 Techniques d'estimation de la pose humaine en 2D

A la suite de notre évaluation de méthodes d'estimation 3D, nous avons émis l'hypothèse que l'utilisation de boîtes de délimitations pourrait améliorer l'estimation 3D. De ce fait nous allons tester des méthodes d'estimation 2D pour comparer leur performance afin de fournir un support à l'estimation 3D.

3.3.1 Estimation de la pose par région

Regional Multi-Person Pose Estimation (RMPE) ou AlphaPose est une méthode d'estimation descendante proposée par Fang et al. [19] pour résoudre les scénarios de pose complexe. Ils proposent la résolution des mauvaises estimations à travers la redéfinition de *l'estimateur de pose unique* (Single-Person Pose Estimator – SPPE) employé. Un nouveau réseau de transformateurs spatiaux symétriques qui est attaché au SPPE pour extraire une région individuelle de haute qualité à partir d'un cadre de délimitation inexact. Une nouvelle branche SPPE parallèle est introduite pour optimiser ce réseau. Une pose paramétrique est introduite pour éliminer les poses redondantes en utilisant une nouvelle métrique de distance de pose pour comparer la similitude de pose. Pour terminer, un nouveau générateur de proposition humaine guidé par la pose (Pose-Guided Proposals Generator – PGPG) est utilisé.

Nous avons testé cette méthode en mettant en place la même implémentation que celle utilisée dans leurs travaux. Nous avons utilisé un réseau VGG [77] comme détecteur humain, un réseau de sabliers empilés comme estimateur de pose pour une seule personne (SPPE), et un réseau de neurones de type ResNet [6] à 18 couches (ResNet-18) comme notre réseau de localisation. En ce qui concerne le SPPE en parallèle, nous avons aussi utilisé un réseau de sabliers empilés à 4 couches.

Nous avons constaté une précision moyenne de score de 61,8 comme l'indique le tableau 2 sur le jeu de données COCO. L'évaluation qualitative sur le même jeu de données et sur des images arbitraires a montré des résultats satisfaisants.

3.3.2 PersonLab

Présentée par Papandreou et al. [29], PersonLab est une méthode d'estimation de la pose ascendante qui propose l'estimation de la pose et la segmentation. Ces fonctions font de cette méthode une approche multitâches. PersonLab propose la détection et association simultanée des points clés du corps dans une architecture multibranche.

Le modèle fonctionne sur une base ResNet pour prédire de manière synchrone les cartes thermiques conjointes de tous les points clés pour chaque personne et leurs déplacements relatifs. Ensuite, le regroupement commence à partir de la détection la plus sûre avec un processus de décodage gourmand basé sur un graphe fondé sur un modèle de squelette.

Nous avons constaté un déphasage entre la performance sur le jeu de donnée COCO et des images arbitraires soumis pour évaluation. Nous avons conclu que cette méthode a été implémentée plus pour l'optimisation des résultats quantitatifs sur le jeu de données COCO que pour un rendement généralisé. Cette méthode a obtenu 68,7 de AP sur le jeu de données COCO et des résultats qualitatifs sur des images arbitraires similaires à ceux de la méthode précédente.

3.3.3 Réseau de sabliers empilés (Stacked Hourglass Network)

Cette méthode proposée par Newell et al.[11] réalise une estimation de personne unique par détection des parties du corps. Le modèle abouti à une carte thermique pour chaque point clé du corps d'une personne cible. Ensuite, le pixel avec l'activation de carte thermique la plus élevée est utilisé comme emplacement prévu pour ce point clé. Le réseau est conçu pour consolider les caractéristiques qui servent à capturer des informations sur la structure complète du corps tout en préservant les détails pour une localisation précise.

Pour évaluer la performance de cette méthode, nous avons utilisé une architecture composée de 8 sabliers (tout comme les auteurs dans l'énonciation de la méthode). Chaque sablier est constitué de : couches convolution et de pooling maximal utilisées pour traiter les caractéristiques jusqu'à une très basse résolution. À chaque étape de max pooling, le réseau se divise et applique plus de convolutions à la résolution d'avant le pooling max. Suivant la résolution, après obtention des caractéristiques correspondant à la résolution la plus basse, le réseau démarre le

suréchantillonnage (up-sampling) et combine progressivement les informations des caractéristiques des différentes échelles. Nous avons constaté une précision moyenne d'environ 63 comme l'indique le tableau 2 sur les données COCO, cependant, nous avons constaté une performance en déclin sur les images avec forte occlusion.

3.3.4 DeepCut

Pishulin et al. [23] ont proposé DeepCut sous la forme d'une méthode ascendante pour l'estimation de la pose.

Bien que cette méthode soit l'une des premières approches de l'estimation ascendante en deux phases, elle demeure assez robuste face aux approches plus récentes et permet des résultats satisfaisants. Nous avons implémenté la méthode conformément aux spécifications utilisées par les auteurs. Nous avons aussi choisi de conserver Fast R-CNN comme détecteur de parties du corps.

En ce qui concerne l'évaluation de la méthode, nous avons constaté que sur le jeu de données, la performance était vraiment moyenne comme l'indique le tableau 2, mais sur des images arbitraires fournies à l'estimation, les performances étaient moins bonnes. La méthode ne fournissait pas des estimations satisfaisantes sur les images avec troncature pour en présence d'occlusion de pixel (Figure 4).

3.3.5 Estimation par affinité des parties

Cette méthode [25] permet de regrouper les parties du corps humain à l'aide des champs d'affinité de partie (PAF). C'est aussi une méthode non paramétrique, pour obtenir un modèle d'estimation de pose de personnes multiples de type ascendant.

Nous avons évalué cette méthode en utilisant le modèle mis à disposition par les auteurs de cette approche. Ce modèle a été initialisé avec 10 couches de VGG-19. Nous avons constaté le rendement de cette approche sur le jeu de données COCO : 62 de précision moyenne (Tableau 2). De plus, tout comme avec les méthodes précédemment examinées, nous avons soumis des images arbitraires au modèle pour évaluer les résultats de manière qualitative (Figure 4). Nous avons constaté que la méthode généralisait bien sur les images arbitraires. Par rapport aux

méthodes précédemment testées, nous avons constaté que les occlusions de pixels étaient mieux gérées par la méthode. Les troncatures n'impactaient pas autant l'estimation.

3.3.6 Synthèse des méthodes 2D évaluées

Nous avons évalué cinq méthodes d'estimation de la pose 2D et relevons des points forts tout comme des faiblesses pour chacune de ces méthodes. Nous présentons un résumé de nos observations dans le Tableau 2 et présentons des exemples sur des images arbitraires sur la Figure 4. Nous avons commencé notre évaluation par la méthode RMPE encore appelée AlphaPose. Nous constatons que cette méthode donne des résultats quantitatifs moyens sur le jeu de données COCO, cependant les résultats qualitatifs sur des images arbitraires présentent des insuffisances sur les images comportant des troncatures (Figure 4). La deuxième méthode évaluée, PerconLab se montre la plus satisfaisante des méthodes sur le jeu de données COCO par opposition la méthode DeepCut qui fournit une précision moyenne la moins impressionnante sur le même jeu de données. Malgré cette performance de PersonLab, nous jugeons la méthode inappropriée du fait de son incapacité à généraliser sur les images arbitraires à forte résolution. La méthode d'évaluation par affinité de parties se montre plus intéressante car nous observons une consistance entre les résultats quantitatifs sur le jeu de données COCO et sur les images arbitraires. Cette méthode présente une résilience par rapport aux autres méthodes en présence de troncature comme le montre la Figure 4.

Tableau 2. – Performance des méthodes 2D évaluées

Méthode	**Type de méthode**	**Avantages**	**Inconvénients**	**Score AP**
RMPE	Estimation de personnes multiples – descendante	Performance satisfaisante sur le jeu de données COCO, Robuste	Temps de calcul plus long lié à l'aspect descendante de la méthode	61,8

PersonLab	Estimation de personnes multiples – ascendante	Meilleure performance sur le jeu de données	La méthode ne généralise pas bien sur les images arbitraires	68,7
Sabliers Empilés	Estimation personne unique basée sur la détection	Bonne performance sur le jeu de données COCO	Inconsistant en présence d'occlusions	66
DeepCut	Estimation de personnes multiples – ascendante	Temps de calcul court	Inconsistant en présence d'occlusions et de troncature d'image	55
Realtime multi-person 2D pose estimation using part affinity fields	Estimation de personnes multiples – ascendante	Architecture robuste et bonne performance sur le jeu de données	Réseau de support gourmand	62

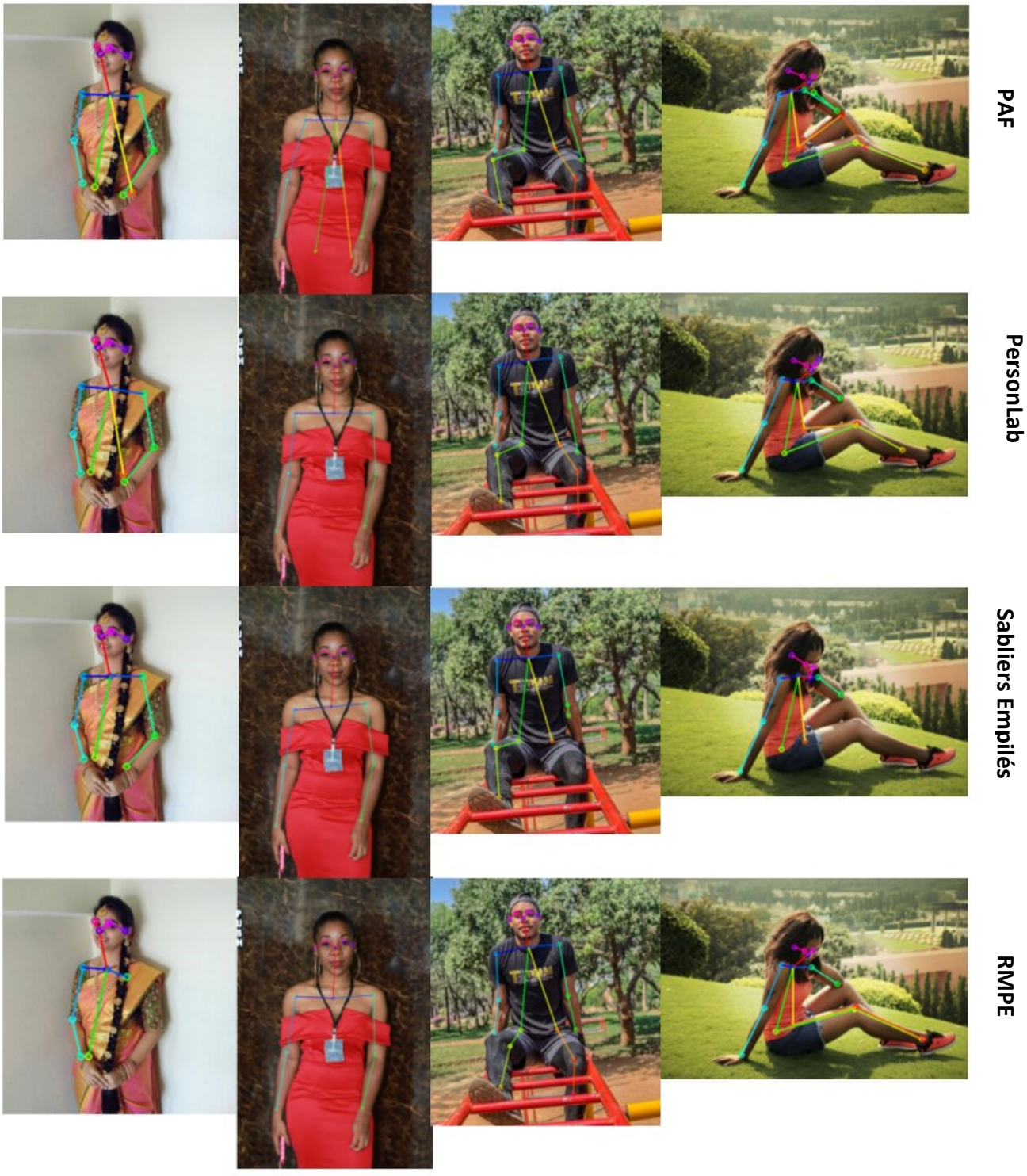

Figure 4. – Comparaison visuelle d'estimation selon différentes approches

3.4 Points sur les résultats de notre observation

À travers cette évaluation, nous avons déterminé que les méthodes pour l'estimation de la pose 3D présentaient des résultats variables sur les images arbitraires en dépit des performances acceptables (des erreurs de construction inférieures à 100 millimètres) sur le jeu de donnes d'évaluation. Nous avons constaté que l'approche de Martinez et al. met en avant la rapidité de l'estimation au détriment de l'exactitude de la pose. Cette méthode reposant exclusivement sur une projection de pose 2D en 3D, la méthode est fortement dépendante de l'approche d'estimation 2D utilisée. Nous avons constaté que la méthode proposée par Kanazawa et al. [61] est meilleure, non seulement car elle permet de régresser la pose 3D sans perte d'information sur l'image, mais aussi parce qu'elle permet d'obtenir bien plus que la pose. Malgré de bonnes performances sur les images arbitraires, la meilleure méthode (Kanazawa et al.) produit des estimations fausses lorsque l'individu sur l'image ne se situe pas au centre de l'image. Nous avons évalué des méthodes d'estimation 2D afin de construire des boîtes de délimitations sur les images. La détermination des boîtes de délimitations exactes dépend de l'exactitude de l'estimation de la pose 2D. À travers notre évaluation, nous avons constaté que la méthode proposée par Cao et al. [25] représentait la meilleure approche. En effet cette méthode a donné une précision moyenne satisfaisante sur le jeu de donnée COCO et bien que certaines méthodes aient donné des précisions plus élevées sur ce jeu de donnée (66 pour les sabliers empilés [11] et 68,7 pour personLab[29]), cette méthode généralise mieux sur les images arbitraires que ces méthodes. Cette méthode est résistante aux troncatures, cela présente un avantage considérable dans la détermination des boîtes de délimitations. Grâce à notre évaluation des méthodes et la détermination des meilleures méthodes, nous avons pu formuler notre approche pour obtenir une estimation de la pose 3D.

Chapitre 4 – Méthode

L'évaluation des méthodes existantes nous a permis de déterminer que l'approche de reconstruction du maillage humain proposée par Kanazawa et al. était convenable pour la réalisation de l'extraction de comportements. De ce fait, nous proposons une architecture. Nous proposons une architecture qui permet l'extraction des points clés 3D par l'articulation de deux méthodes existantes. À travers nos différentes évaluations, nous avons constaté que les méthodes aussi résistantes qu'elles soient échouent parfois dans l'estimation de la pose, car formulée pour réussir sur des images présentant des individus au centre de l'image et avec des résolutions précises (224x224 pour des résultats optimaux). De ce fait, quand certaines images ne respectent pas les propositions de ces méthodes, elles obtiennent des résultats moyens. **Nous suggérons une estimation de la pose 3D en deux phases** (annexe A). Dans la première phase, nous estimons et entreposons les points clés 2D en utilisant la méthode des champs d'affinité de parties [25]. Puis dans la seconde phase, nous procédons à l'estimation et au stockage des points clés en nous appuyant sur la méthode reconstruction de maillage [61]. L'ensemble de ces deux phases décrit notre approche (Figure 5).

Dans ce chapitre, nous présentons la première phase de la méthode, la seconde puis les différents réseaux de support utilisés lors de nos expérimentations afin d'obtenir une approche fournissant de meilleurs résultats.

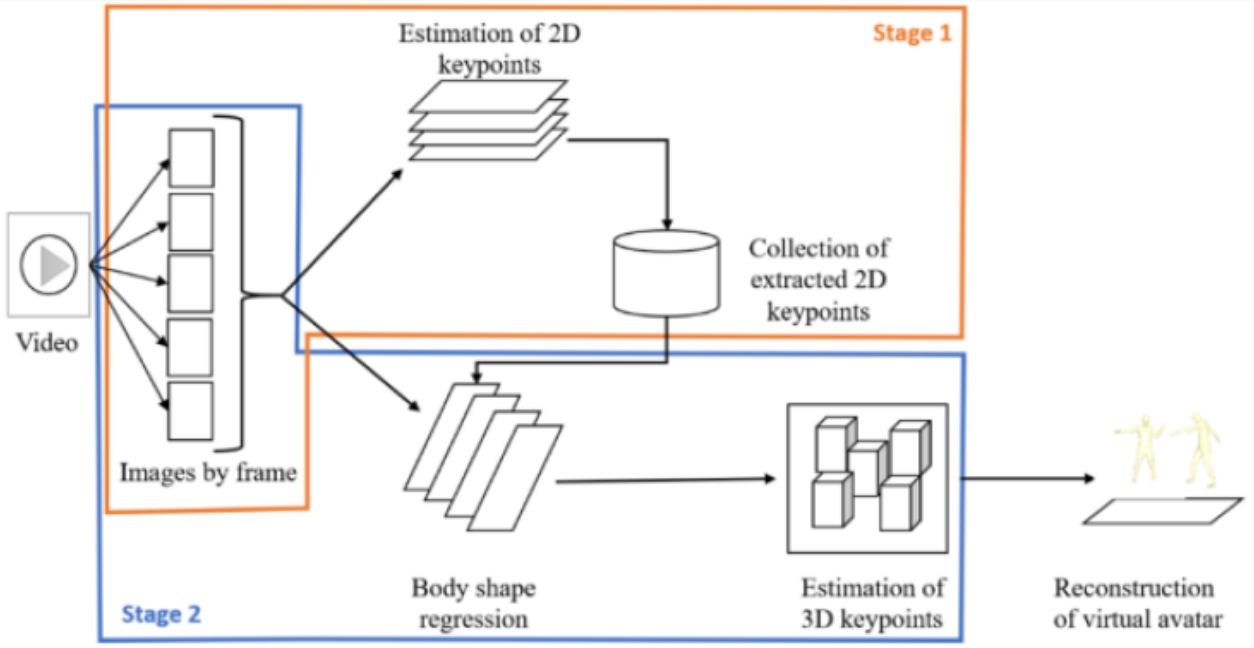

Figure 5. – Architecture globale (annexe A)

4.1. Première phase

Dans notre architecture, l'emploi d'une méthode d'estimation 2D permet de traiter les images en entrée afin de déterminer les boîtes de délimitations plus exactes sur les images. Le but de ces boîtes étant de délimiter la portion d'une image où se situe un individu. La détermination des boîtes de délimitations intervient dans le traitement de l'image.

Nous nous appuyons sur les champs d'affinité [25] pour réaliser cette phase compte tenu de ces performances lors de notre évaluation des méthodes existante (Tableau 2).

Cette méthode présente une approche ascendante pour l'estimation de la pose de personnes multiples dans une image et produit, en sortie, les emplacements 2D des points clés pour chaque personne dans l'image, sans utiliser un détecteur de personne. Le modèle définit une architecture de réseau qui prédit de manière itérative des champs d'affinité qui encodent des cartes de confiance d'association et de détection de partie à partie.

Le réseau est divisé en deux branches pour chaque étape de l'estimation (Figure 6) : une première qui prédit les cartes de confiance et la seconde qui prédit les champs d'affinité. Chaque section

est une architecture de prédiction itérative, qui affine les prédictions à travers plusieurs étapes, avec une supervision intermédiaire à chaque étape.

Premièrement, une image en entrée est soumise à un réseau à convolution pour générer les cartes de caractéristiques F (partie verte de la Figure 6) qui seront ensuite soumises à la première étape (étape t = 1). Lors de cette première étape, ces cartes de caractéristiques sont envoyées à deux branches de convolutions. La première branche permet de produire un ensemble de cartes de confiance de détection $S^1 = \rho^1(\boldsymbol{F})$ (3) et la seconde branche un ensemble de champs d'affinité de pièce $L^1 = \phi^1(\boldsymbol{F})$ (4), où ρ^1et ϕ^1 représentent les convolutions à la première étape (couleur bleu Figure 6). Ensuite, l'ensemble des champs d'affinité de pièces et l'ensemble de cartes de confiance prédites par les deux branches de la première étape sont concaténés avec la carte de caractéristiques de l'image $\boldsymbol{F}$ pour raffiner les prédictions pour les étapes suivantes (couleur jaune Figure 6) :

$$S^t = \rho^t(\boldsymbol{F}, S^{t-1}, L^{t-1}), t \geq 2 \qquad (5)$$

$$L^t = \phi^t(\boldsymbol{F}, S^{t-1}, L^{t-1}), t \geq 2 \qquad (6)$$

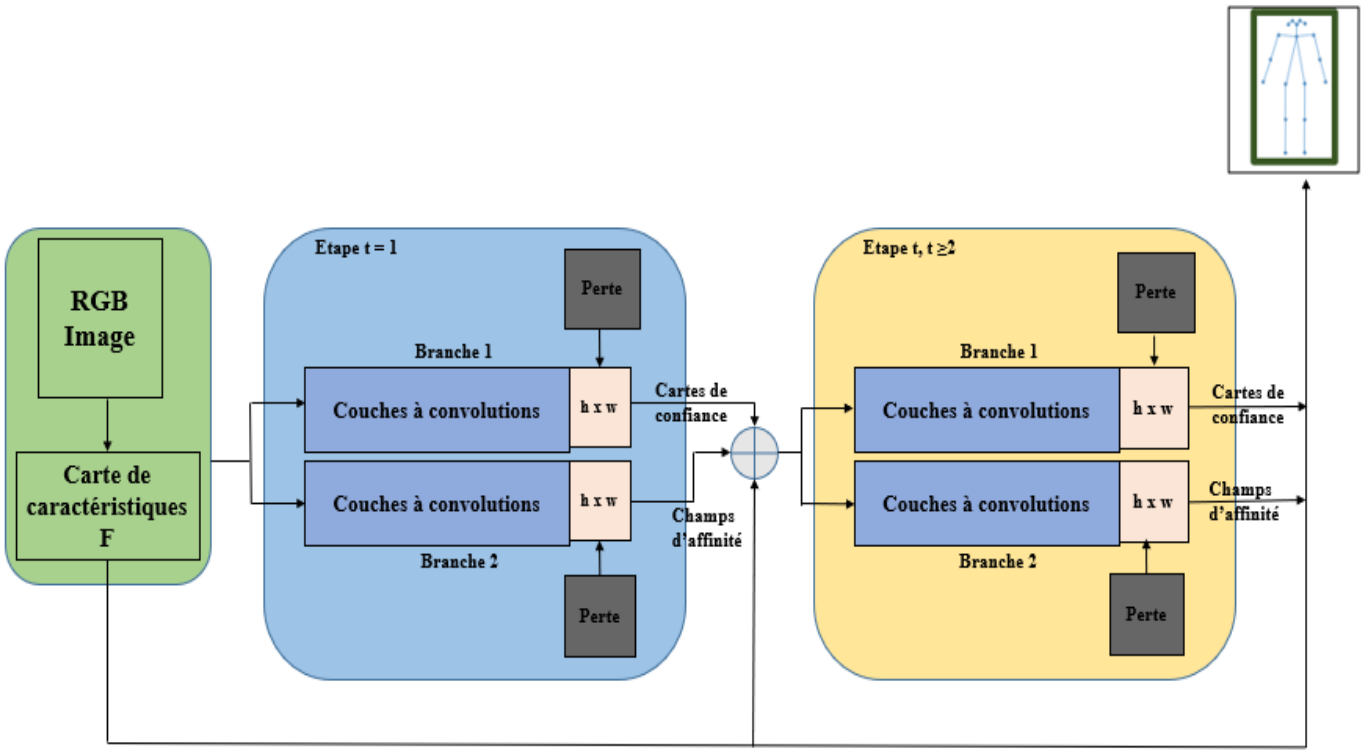

Figure 6. – Architecture de la première phase

4.2. Deuxième phase

Nous utilisons, les résultats de la première phase (prédiction de la pose 2D) pour supporter l'encodage de l'image pour la réalisation de la régression de pose 3D par le modèle de Kanazawa et al. [61]. L'utilité des résultats obtenus dans la première est la détermination plus précise des boîtes de délimitations. Avec les détections de coordonnées 2D de la première phase, nous encodons les parties de l'image avec plus d'assurance que nous ne perdons pas des parties de l'image contenant des données pertinentes. L'approche utilisée génère non seulement les coordonnées des points clés en 3D, mais aussi un maillage humain 3D à partir d'image. Avec cette approche, le maillage 3D d'un corps humain est encodé à l'aide de SMPL qui génère des moulages de corps humains en fonction de la variation de la taille, du poids et de la proportion du corps, et de la déformation de la surface due aux mouvements. Il utilise 10 coefficients d'un ensemble d'analyse en composantes principales (PCA) pour ajuster la forme tandis que la pose est modélisée par la rotation et l'ensemble des points clés.

L'objectif est d'obtenir des paramètres $\Theta = (\beta, \theta, \boldsymbol{R}, t, s)$ pour chaque image. $\beta \in \mathbb{R}^{10}$ représente la forme du corps, $\theta \in \mathbb{R}^{3\boldsymbol{K}}$ la rotation 3D relative de $\boldsymbol{K}$ = 23 points, $\boldsymbol{R} \in \mathbb{R}^{\boldsymbol{3\times 3}}$ rotation globale, $t \in \mathbb{R}^2$ la transposition et $s \in \mathbb{R}$ L'échelle. Le modèle comporte un module de régression qui extrait les paramètres $\boldsymbol{\Theta}$ à partir des caractéristiques d'une image. Les caractéristiques et les paramètres SMPL initiaux (β, θ) sont concaténés et alimentés au module de régression (Figure 7).

Avec SMPL un maillage triangulé avec N = 6980 sommets est généré par la fonction $M(\theta, \beta) \in \mathbb{R}^{3\times N}$. Les points clés 3D utilisés pour l'erreur de reprojection, $X(\theta, \beta) \in \mathbb{R}^{3\times p}$, sont obtenus par régression linéaire à partir des sommets du maillage final. Une projection orthogonale Π est utilisé pour le calcul de la projection $X(\theta, \beta)$ des points clés 3D sur un plan 2D suivant la formule

$$\hat{x} = s\Pi\big(\boldsymbol{R}\, X(\theta, \beta)\big) + t \qquad (7)$$

L_{rep} représente la pénalité pour minimiser l'erreur de projection. Cette pénalité est calculée suivant la formule :

$L_{rep} = \sum \| v_i(x_i - \hat{x}_i) \|_1$ (8) ou $x_i \in \mathbb{R}^{2\times K}$ vérité terrain numéro i des points clés 2D et $v_i \in \{0,1\}^K$ la visibilité (1 si visible, 0 sinon) pour chacun des $\boldsymbol{K}$ points.

Avec cette approche, le réseau infère les paramètres de maillage 3D et la caméra de manière réaliste, un réseau discriminant est utilisé pour déterminer si les paramètres 3D correspondent à des corps d'humains réels ou non. Ce discriminateur agit comme une supervision et permet au réseau d'apprendre l'angle réaliste de chaque point clé dans le but de trouver l'emplacement approprié des parties du corps et d'empêcher la reconstruction de formes corporelles inhabituelles.

Notre approche utilise non seulement l'image, mais également les points clés détectés par la première phase pour le calcul des boîtes de délimitations pour un meilleur prétraitement de l'image par la régression de maillage 3D. Cette amélioration aide au recadrage. Le recadrage est effectué en fonction du centre de la boîte de délimitations et l'échelle. Le recadrage est de forme rectangulaire marquée par la valeur la plus élevée entre la hauteur et la largeur du cadre de délimitation (déterminé par les coordonnées 2D estimées lors de la première étape).

Avec l'intervention des points clés 2D, le processus de reconstruction du maillage ressemble à celui décrit dans la figure suivante

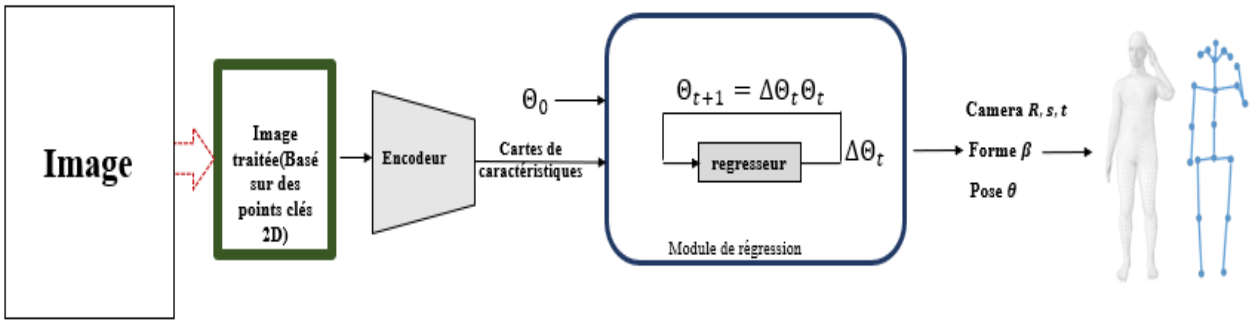

Figure 7. – Description de la deuxième phase

Le succès d'une méthode d'estimation (régression) de la pose reposant majoritairement sur le réseau de support, nous avons considéré plusieurs réseaux de support pour articuler notre approche. Les méthodes choisies sont : AlexNet, MobileNet, ResNet et VGG. Dans la suite, nous présentons de manière succincte ces réseaux.

4.3. Réseaux de support (Backbones)

4.3.1. AlexNet

Ce réseau a été introduit par Krizhevsky et al. [1]. Il était considéré comme référence pendant de nombreuses années en matière de classification d'images et de segmentation d'objets. AlexNet se compose de 5 couches à convolutions et de 3 couches entièrement connectées, et c'était le premier réseau à utiliser la fonction d'activation ReLu au lieu du tanh et à obtenir de meilleurs résultats six fois plus rapidement. Cette architecture a introduit une nouveauté, des mises à l'échelle (pooling) maximales qui se chevauchent. Les deux premières couches à convolution utilisaient 96 noyaux de taille 11 x 11 x 3 et sont liées au pooling maximal qui se chevauchent. Ces gros noyaux ont aidé à extraire des fonctionnalités précieuses d'une image d'entrée. De plus, pour éviter le sur apprentissage, les auteurs ont utilisé la technique d'abandon (dropout) dans les couches entièrement connectées ainsi que des augmentations telles que le retournement et la

rotation aléatoires. Dans nos expériences, nous avons utilisé un réseau entraîné sur ImageNet [78].

4.3.2. MobileNet

MobileNet-v2 [79] est un réseau neuronal à convolution de 53 couches de profondeur. MobileNetV2 est une amélioration par rapport à MobileNetV1 [80] en introduisant des résidus inversés et des goulots d'étranglement linéaires. La convolution en profondeur ne peut pas modifier le nombre de canaux, ce qui entraîne une extraction de caractéristiques limitée par le nombre de canaux d'entrée. Le résidu inversé résout ce problème. La fonction d'activation RELU6 dans le bloc résiduel inversé accélère l'apprentissage, supprime la disparition du gradient et augmente la stabilité du modèle. Cependant, les transformations ReLU peuvent entraîner une perte importante d'informations sur les caractéristiques de faible dimension, de ce fait, MobileNetV2 remplace la fonction ReLU6 dans la dernière couche du résidu inversé par une fonction d'activation linéaire pour réduire la perte d'informations.

4.3.3. ResNet

ResNet [6] est un réseau résiduel profond, une sous-classe de réseaux neuronaux à convolution qui sont très populaires sur les tâches de classification d'images. ResNet parvient à surmonter le problème de disparition du gradient qui se produit lors de la formation d'architectures profondes en établissant la notion de « saut de connexions ». Ces « raccourcis » sautent généralement par-dessus deux ou trois couches, qui sont généralement des couches à convolutions contenant l'activation ReLU et la normalisation par lots. Il crée une connexion qui ajoute l'entrée à la sortie des couches et qui permet aux dégradés de s'écouler pendant l'entraînement. Le saut de couches simplifie également efficacement le modèle et facilite l'entraînement en utilisant moins de couches au début de la formation et en augmentant progressivement leur nombre.

4.3.4. VGG

Cette architecture provient du groupe de géométrie visuelle (VGG) [77]. Il hérite de l'architecture AlexNet, mais remplace les grands filtres de la taille d'un noyau qui étaient utilisés avec plusieurs filtres empilés de la taille d'un noyau 3 x 3. Il permet au réseau d'aller plus loin et aide à conserver

les propriétés de niveau plus fin des images. VGG a également introduit le concept de blocs de couches, qui est devenu un thème commun dans les réseaux suivants. Dans nos expériences, nous avons utilisé VGG 16 et VGG 19 entraîné sur ImageNet. Le « 16 » et « 19 » font référence au nombre de couches.

Nous avons formulé une méthode qui combine plusieurs approches pour obtenir une solution stable pour l'estimation de la pose tridimensionnelle. Nous avons utilisé des prédictions de la pose bidimensionnelle obtenue par une approche multiétages [25] qui délivre un ensemble de cartes de confiance de détection et de champs d'affinité partielle passés par l'inférence gloutonne pour avoir les points clés 2D pour chaque individu dans l'image, appelée correspondance bipartite. Par la suite, nous avons exploité les ces prédictions obtenues pour raffiner le traitement des images afin de les soumettre à une approche de régression de la pose tridimensionnelle.

Dans l'optique de consolider notre méthode, nous avons utilisé des réseaux de supports variés pour supporter notre approche. Dans le chapitre suivant, nous présenterons les résultats que nous avons obtenus ainsi que les réseaux de supports les plus adaptés en nous basant sur ces résultats obtenus.

Chapitre 5 — Résultats extraits

Dans ce chapitre, nous explorons les expérimentations effectuées sur notre méthode. Nous comparons les résultats obtenus en utilisant différents réseaux de support en combinaison avec les méthodes choisies respectivement pour la première et pour la seconde phase. Nous confrontons nos résultats aux méthodes initiales sur les jeux de données COCO pour l'estimation 2D et Human3.6M pour l'estimation finale de la pose 3D.

5.1. Estimation 2D

Nous avons conduit nos expérimentations en utilisant le Framework Tensorflow [81] sur une machine GeForce RTX 2070 Super. Pour les entrainements, nous avons considéré un lot (batch size) de 10, 55 époques (epoch) et un taux d'apprentissage de 4e-5. Nous avons expérimenté la méthode avec plusieurs réseaux de support. Dans nos expériences, nous avons utilisé 6 étapes (t=6). Pour commencer, nous avons utilisé un réseau VGG-19. Dans cette architecture, nous fournissons une image de couleur au réseau de support (les 10 premières couches de VGG-19) pour produire l'ensemble de cartes d'entités F.

Les deux premières couches à convolutions filtrent l'image d'entrée avec 64 noyaux de taille 3×3 avec une foulée de 1×1, suivis d'un pooling maximal 2×2 avec une foulée de 2×2. Ensuite, deux couches à convolution avec 128 noyaux de taille 3×3 avec une foulée de 1×1, suivi d'un pooling max 2×2 avec un pas de 2×2. Les quatre troisièmes couches à convolution filtrent avec 256 noyaux avec une taille de 3×3 avec une foulée de 1×1, suivis par un pooling maximal 2×2 avec une foulée de 2×2. Ensuite, trois couches à convolution filtrent avec 512 noyaux de taille 3×3 avec une foulée de 1×1, et la dernière couche filtrante convolutive avec 128 noyaux de taille 3x3 avec une foulée de 1×1.

Ensuite, les cartes d'entités F sont utilisées pour le module d'association de poses pour générer des cartes de confiance S et des cartes associées aux parties (champs d'affinité des parties) L. Il y a deux étapes dans le module d'association de poses: (S1, L1) et (St, Lt). La première étape produit (S1, L1), où ρ^1 est composé de trois couches à convolution avec 128 noyaux de filtre de taille 3×3

avec un pas de 1×1, une couche à convolution avec 512 noyaux de filtre de taille 1 × 1 avec une foulée de 1×1 et une couche a convolution avec 19 (nombre de points clés) noyaux de filtrage de taille 1×1 avec une foulée de 1×1 pour S1. Φ^1 et ρ^1 utilise les mêmes couches à convolutions à l'exception de la dernière couche à convolution dans laquelle se trouve une couche à convolution avec 38 (nombre de points clés x 2) noyaux de filtrage de taille 1×1 avec une foulée de 1 × 1 pour L1.

À partir de la deuxième étape produisant (St, Lt), ρ^t est composé de cinq couches à convolutions avec 128 noyaux de filtre de taille 7x7 avec une foulée de 1×1, et une couche à convolutions avec 19 noyaux de filtrage de taille 1×1 avec une foulée de 1×1 pour St. Φ^t et ρ^t utilise les mêmes couches à convolutions à l'exception de la dernière couche à convolutions dans laquelle se compose d'une couche à convolutions avec 38(nombre de points clés x 2) noyaux de filtrage de taille 1×1 avec une foulée de 1×1 pour Lt.

L'utilisation de ce réseau donne un caractère robuste à la méthode et permet d'obtenir de bons résultats. Néanmoins, VGG est un réseau assez lourd. Nous avons donc expérimenté cette méthode en remplaçant ce réseau de support par d'autres, jugés plus légers afin de mesurer le rendement. Sur le réseau de support Mobilenet v2, nous avons testé la même implémentation avec 12 couches à convolutions comme couches d'extraction de caractéristiques.

Sur le réseau de support Resnet-50, nous n'utilisons qu'une seule couche à convolutions et deux blocs résiduels de Resnet-50.

Les blocs résiduels de Resnet-50 se composent de 22 couches à convolutions. La première couche à convolutions filtre l'image d'entrée avec 64 noyaux de taille 7x7 avec une foulée de 2×2, suivie d'un pooling maximal de 3×3 avec une foulée de 2×2. Le premier bloc résiduel se compose d'un bloc convolutif et de deux blocs d'identification qui filtrent les noyaux [64, 64, 128] de taille [1×1, 3×3, 1×1], et le deuxième bloc résiduel est constitué d'un bloc convolutif et trois identifiants des blocs qui filtrent [128, 128, 512] noyaux de taille [1×1, 3×3, 1×1].

Ensuite, les cartes de caractéristiques F sont utilisées comme entrées pour générer des cartes de confiance S et des cartes associées à des parties L. Le réseau du module de couplage de poses pour générer la première étape (S1, L1) et la deuxième étape (St, Lt).

Nous avons mesuré la précision moyenne de chacun des réseaux afin de mieux comparer les résultats. Nous avons aussi évalué les résultats qualitatifs sur des images arbitraires.

En dépit de l'inconvénient majeur en temps de calcul et ressource que le réseau VGG présent, nous avons observé la précision la plus élevée en l'utilisant. Cependant, le réseau ResNet s'est avéré plus léger et plus rapide. De plus, nous avons reporté une précision qui avoisine celle de VGG et un rappel moyen plus élevé que ce dernier (comme le montre le Tableau 3 ci-dessous).

Tableau 3. – Comparaison des performances des réseaux de support 2D

Réseaux	AP	AP^{50}	AP^{75}	AR
VGG	**62**	**84,9**	**67,4**	66,5
MobileNet	57,2	79,7	61,6	60,9
Resnet50	59,2	84,5	64,7	**68,8**

Dans l'approche que nous proposons, les résultats de la première phase sont essentiels pour améliorer les estimations 3D. De ce fait, nous avons plus considéré le poids au réseau de support obtenant une précision moyenne plus grande en dépit de sa complexité en temps de calcul. Nous avons décidé de privilégier VGG. Cependant, nous avons aussi observé que dans plusieurs cadres ResNet peut se révéler être un meilleur choix, car alliant légèreté et performance. Nous montrons des exemples d'estimation de points clés sur la figure ci-dessous (Figure 8). Nous avons aussi constaté que, dans certains cas d'images avec troncature, le réseau ResNet – 50 a permis d'estimer des points clés que le réseau VGG 19.

VGG 19 Mobilenetv2 ResNet -50

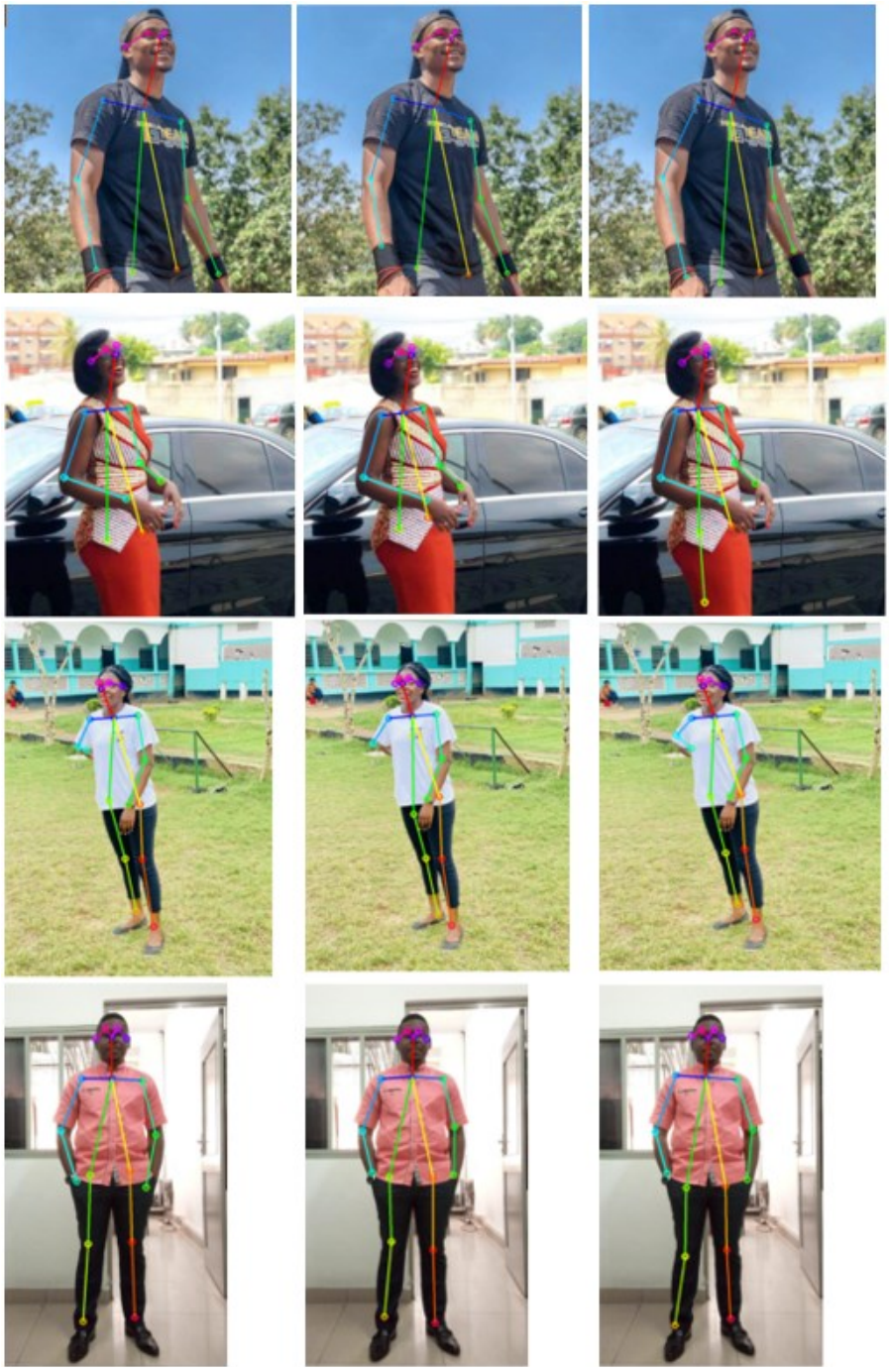

Figure 8. – Exemple d'estimation 2D selon les différents réseaux de support

5.2. Estimation 3D

Nous avons conduit nos expérimentations en utilisant le Framework Tensorflow sur une machine GeForce RTX 2070 Super. Les réseaux de support ont été pré-entraînés sur ImageNet. Toutes les couches utilisent des activations ReLU sauf la couche finale, où la sortie produit des caractéristiques $\Phi \in \mathbb{R}^{2048}$. Le module de régression 3D se compose de deux couches entièrement connectées avec 1024 neurones chacune avec une couche d'abandon entre les deux, suivies d'une couche dense de 85 neurones. Il y a eu 3 itérations au total pour la procédure de régression pour chacun des réseaux de support choisis (à savoir ResNet 50, VGG-19 et AlexNet). Le discriminateur est constitué de deux couches entièrement connectées à 10 et 5 neurones puis 1 neurone en sortie. Premièrement, la pose est convertie K matrices de dimension 3 par 3 chacune (matrice de rotation) en utilisant la formule de Rodrigues[82]. K correspond au nombre de points clés régressés et est égale à 23. Ensuite, chaque matrice de rotation est envoyée à un réseau d'intégration commun de deux couches entièrement connectées avec 32 neurones cachés.

Le discriminateur pour la distribution globale des poses concatène toutes les représentations ***A*** x 32 à travers deux autres couches entièrement connectées de 1024 neurones chacune et délivre un résultat unidimensionnel. Nous avons fixé les taux d'apprentissage du codeur et du réseau discriminateur respectivement à 1×10^{-5} et 1×10^{-4}. Nous utilisons le solveur Adam et nous entraînons pendant 50 époques (epoch) avec une taille de lot (batch size) de 64. Ce choix du nombre d'époques se justifie par l'observation d'un plateau en termes de résultats (précision) au-delà de 50 époques. Concernant la taille de lot, la valeur choisie constitue un choix personnel.

Nous avons observé les performances de la méthode en prenant pour réseaux de support : VGG-16, AlexNet et enfin ResNet-50 (Tableau 4). Ainsi, nous avons observé non seulement les résultats qualitatifs sur des images et vidéos arbitraires, mais aussi les résultats quantitatifs sur le jeu de données Human3.6M. Que ce soit sur les données Human3.6M ou sur les séquences arbitraires, ResNet-50 s'est montré le plus adéquat en raison des résultats obtenus. En effet, nous avons observé une erreur de position moyenne par point clé (MPJPE) de 86,2 millimètres avec ResNet-50 comparé à 93,5 millimètres et approximativement 100 millimètres pour VGG et AlexNet respectivement. Les résultats démontrent que l'erreur de reconstruction est minimale avec le

réseau ResNet -50, mettant ainsi en lumière que l'usage de ce réseau avec la méthode formulée permet une estimation avec moins d'erreurs.

Tableau 4. – Comparaison des performances des réseaux de support 3D

Réseaux	MPJPE	Erreur de Reconstruction
ResNet	**86,2**	59,7
VGG	93,5	**57,6**
AlexNet	100,1	67

La comparaison des résultats des différents réseaux de support avec la méthode choisie a permis de déterminer que le réseau ResNet-50 est le réseau le plus adéquat pour obtenir de meilleures prédictions.

L'usage de ce réseau supporté par les résultats de la première phase nous a permis d'obtenir de meilleurs résultats sur les médias arbitraires. Nous avons soumis des images arbitraires à l'estimation de la pose sans support de coordonnées 2D et à l'estimation selon notre approche qui consiste à estimer les points clés d'une image et à utiliser ces points clés pour les traitements de l'image avant régression de la pose 3D. Cette comparaison nous a permis de constater de meilleures régressions du maillage (utilisé pour représenter la pose 3D), et par conséquent de meilleures estimations de coordonnées. Sur la figure 9, nous pouvons observer les différentes estimations. Les coordonnées 2D sont utilisées pour estimer les boîtes de délimitations (en bleu sur la figure 9) afin de mieux traiter l'image et de donner un meilleur point de départ pour la régression de la pose 3D.

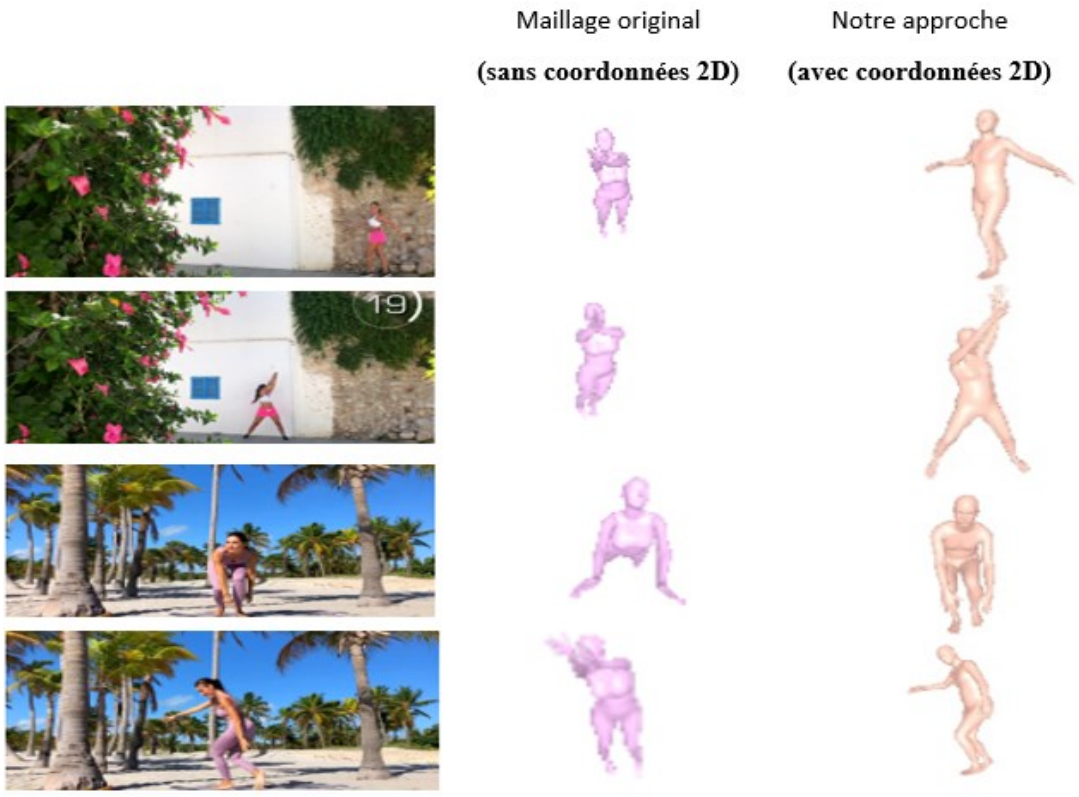

Figure 9. – Visualisation des maillages avec et sans coordonnées 2D

Nous avons pu déterminer que l'usage des coordonnées 2D ajouté à la méthode tel que proposé par Kanazawa et al. permettait d'augmenter les marges de succès du fait du support dans la détermination des boîtes de délimitations.

Nos différentes expérimentations nous ont permis de déterminer les combinaisons nécessaires pour obtenir des résultats pertinents. Premièrement, nous avons entraîné et testé un modèle d'estimation de la pose 2D avec différents réseaux de support. Nous avons observé que le réseau VGG garantissait de bons résultats comparativement au réseau ResNet et MobileNet. Par la suite, nous avons entraîné et testé le système de régression de la pose 3D avec les réseaux de support ResNet, VGG et AlexNet. Si le réseau VGG s'est montré le plus adéquat pour la phase d'estimation de la pose 2D, il n'est pas le réseau qui a permis d'obtenir l'erreur de position moyenne par point clé minimale. Nous avons observé une erreur de position moyenne par point clé minimale avec

le réseau de support ResNet, tandis que le réseau AlexNet en dépit de sa notoriété sur les travaux impliquant l'imagerie, a fourni les résultats les moins impressionnants.

Le but de ces expérimentations a été de déterminer la combinaison qui favorise de bons résultats pour notre approche. En effet nous avons pu dégager que pour notre méthode, le réseau de support VGG était à privilégier pour la première phase (estimation des coordonnées 2D) et que le réseau ResNet devrait être le réseau de support part prédilection pour la régression de la pose 3D finale. La détermination de la bonne combinaison des réseaux de support nous a permis d'entamer la reproduction des comportements à proprement parler. Dans le chapitre suivant, nous présentons notre approche pour la réalisation de cette reproduction.

Chapitre 6 — Reconstruction de mouvements

Les différentes expérimentations réalisées dans le chapitre précédent nous ont permis de déterminer les réseaux de support efficients et assurer l'estimation de la pose 3D telle que présenté dans le chapitre 4. Nous avons estimé la pose 3D en utilisant notre approche qui combine un modèle d'estimation des coordonnés 2D pour déterminer des boîtes de délimitations autour d'un individu à un modèle de régression de la pose 3D. Dans le chapitre précédent, nous avons présenté les résultats des réseaux de support et réussi à établir les bases de notre méthode d'estimation de la pose. Dans le présent chapitre, nous abordons la reconstruction des mouvements décrits par une succession de poses. Nous décrivons principalement les techniques employées pour reproduire par un avatar virtuel, les comportements (ensemble de poses) décrits dans une vidéo. Nous y discutons la structure de stockage des coordonnées estimées, la détermination du type d'avatar pour ce genre de tache et le modèle adopté pour reconstruire les mouvements en fonction des coordonnées 3D.

6.1. Stockage des points clés 3D estimés

Le but de l'estimation de la pose 3D étant de reproduire en environnement virtuel les comportements observés dans des vidéos, la détermination de la structure de stockage pour les informations estimées représente une partie décisive.

Nous avons évalué différentes approches pour entreposer les informations estimées. La nature de nos travaux ne visant pas principalement l'exécution de la reproductibilité en temps réel, nous avons décidé d'entreposer le résultat d'une estimation à travers un fichier. Nous avons comparé le format JSON et CSV pour cette tâche. En se basant sur notre approche qui consiste à évaluer une vidéo comme une série d'images, pour entreposer les coordonnées afin de les reproduire, la méthode la plus optimale est d'entreposer ces coordonnées image par image. Avec cette approche, le format JSON n'est pas le meilleur car étant moins compact que le format CSV. L'utilisation du fichier CVS offre plusieurs avantages dont la lisibilité, la simplicité de création et surtout la consommation moins importante en ressource mémoire contrairement au format

JSON. Ainsi pour les longues vidéos, le fichier final obtenu demeure compact et les données facilement accessibles.

6.2. Adressage des points clés du corps humain

La reproduction des comportements implique que les parties du corps de l'avatar bougent de manière similaire aux parties de l'individu dont les coordonnées de points clés ont été extraites. Dans ce sens, nous utilisons un caractère avatar humanoïde qui offre la possibilité d'accéder à la manipulation de certains points clés du corps.

La première étape de cette tâche de reproduction consiste à établir la relation entre les points clés extraits et les points manipulables de l'avatar. Il existe une légère variation entre les points clés. Les caractères avatars humanoïdes et le squelette extrait offrent tous le même nombre de points clés à manipuler (21). Cependant, les points diffèrent comme le montre la Figure 10. Nous recueillons les coordonnées des yeux et oreilles, points clés dont l'avatar n'offre pas la manipulation. Réciproquement, les caractères avatar disposent de points clés que nous n'extrayons pas. Ces variations entrainent un ajustement.

Nous décrivons l'adressage effectué entre un avatar et l'ossature que nous extrayons sur la figure suivante

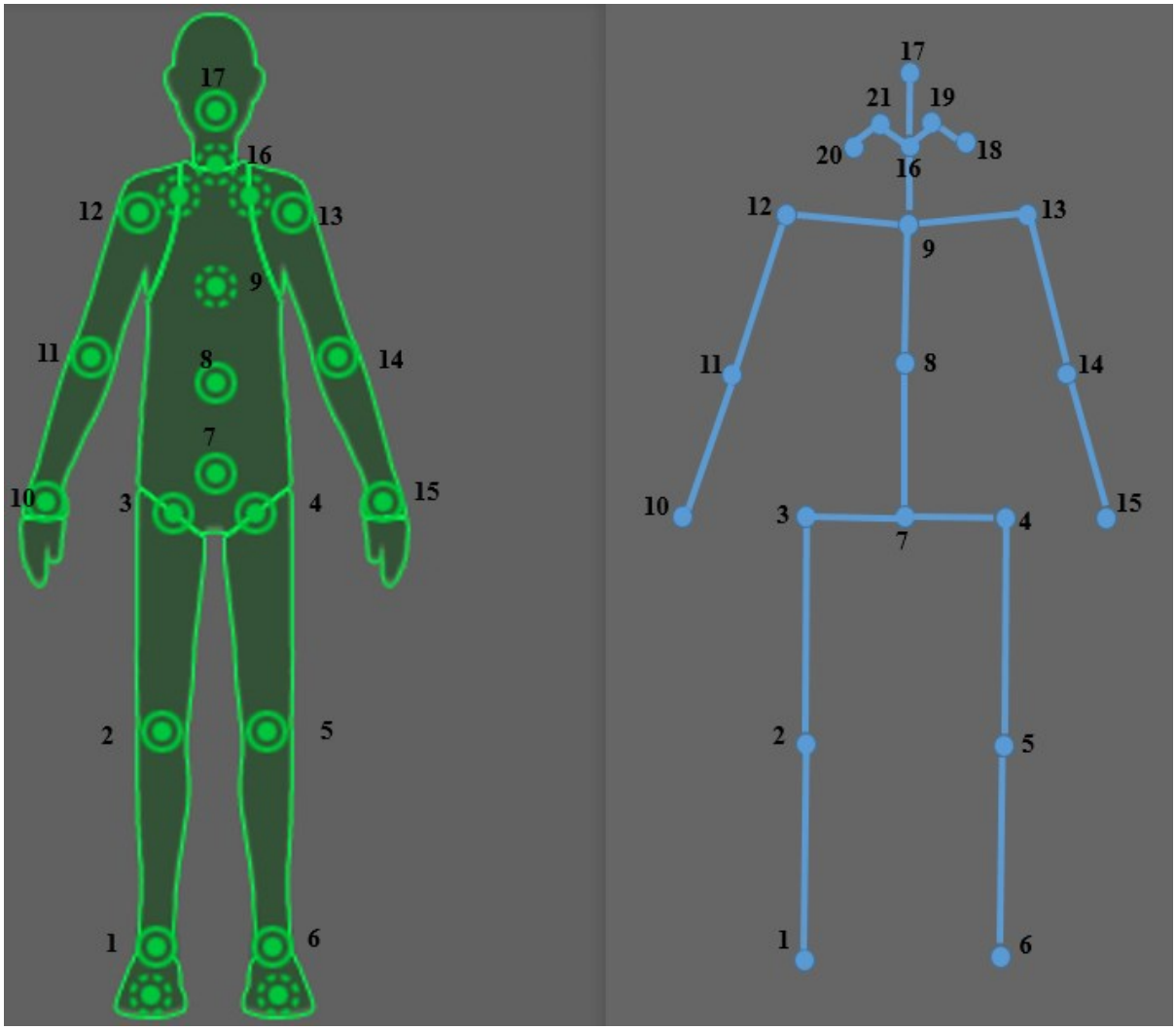

Figure 10. – Adressage des points clés de l'avatar

Nous présentons à gauche le modèle avatar et les points clés disponibles dont nous disposons dans notre environnement virtuel (unity3D) et à droite les points clés extraits. Sur le modèle, les points encerclés de façon discontinue désignent les points facultatifs. L'absence de ces coordonnées pour ces points ne pénalise pas une animation générale de l'avatar.

6.3. Détermination des mouvements

L'adressage des points clés à lui seul ne suffit pas pour reproduire le comportement. L'ensemble des mouvements définit le comportement. La tâche effectuée jusqu'à présent nous a permis d'extraire les positions des points clés tridimensionnelles. Une reproduction de comportement dans notre environnement virtuel passe par la disposition des informations de position, de

rotation et d'orientation de chaque point clé. Nous avons tout d'abord établi une dépendance des points clés. Cette dépendance décrit l'alignement des points clés les uns par rapport aux autres et donc l'influence des coordonnées d'un point clé sur un autre. Après l'établissement des dépendances des points clés, nous adressons la cohésion des mouvements de points clés. En d'autres termes, lors d'un déplacement, ces points clés qui sont en relation effectuent un mouvement harmonique. Ultimement, nous formulons les rotations en considérant les angles que forment les points clés et les points clés adjacents. Le tableau 5 présente la liste des dépendances effectuées.

Tableau 5. – Tableau de hiérarchisation

Points clés primaires		**Points clés adjacents**
Noms	Numérotation	
Pelvis	1	–
Hanche côté droit	2	Genou droit
Genou droit	3	Cheville droite
Cheville droite	4	–
Hanche côté gauche	5	Genou gauche
Genou gauche	6	Cheville gauche
Cheville gauche	7	–
Milieu de dos	8	Cou
Cou	9	Front
Nez	10	–
Front	11	–
Épaule gauche	12	Coude gauche
Coude gauche	13	Poignet gauche
Poignet gauche	14	–
Épaule droite	15	Coude droit
Coude droit	16	Poignet droit
Poignet droit	17	–

6.4. Système de conseils

La reproduction de comportement par un avatar présente une multitude d'applications. Nous avons étendu nos travaux à la réalisation d'un système de conseil basé sur la reproduction de comportement de patients Alzheimer. Nous décrivons le système de conseil comme un système permettant d'observer différents comportements d'un patient atteint de la maladie d'Alzheimer (1) et de choisir l'attitude appropriée selon le comportement du patient (2).

Les comportements propres aux patients Alzheimer sont variés et dans le cadre l'application de nos travaux, nous avons choisi d'adresser un sous-ensemble de ces comportements. Ainsi pour notre système, nous avons choisi les comportements suivants :

- Anxiété
- Colère
- Apathie
- Agitation
- Dépression
- Stress
- Joie

Une vidéo peut décrire plusieurs comportements. En effet afin de permettre la reproduction de plusieurs comportements (en addition des comportements initiaux mentionnés plus haut), nous avons considéré un système permettant de dégager les différentes nuances de comportements. Ainsi, nous avons décidé de construire un **éditeur de comportements**.

Nous fournissons un éditeur de comportements pour permettre de dégager différents comportements afin de mieux capturer les comportements à reproduire dans le système de conseils. Grâce à cet éditeur de comportements, nous permettons de définir les différents intervalles pour chaque comportement dans une vidéo. Pour chaque intervalle (par extension comportements), une estimation de la pose est réalisée. La Figure 11 décrit l'environnement d'étiquetage comportemental d'une vidéo. Le début et la fin du comportement sont délimités et

la confirmation de l'intervalle déclenche l'exécution de l'estimation de la pose pour l'intervalle en question. À la suite de cette estimation, le comportement est rajouté aux comportements dans le système de conseils afin de fournir la reproduction du comportement par un avatar.

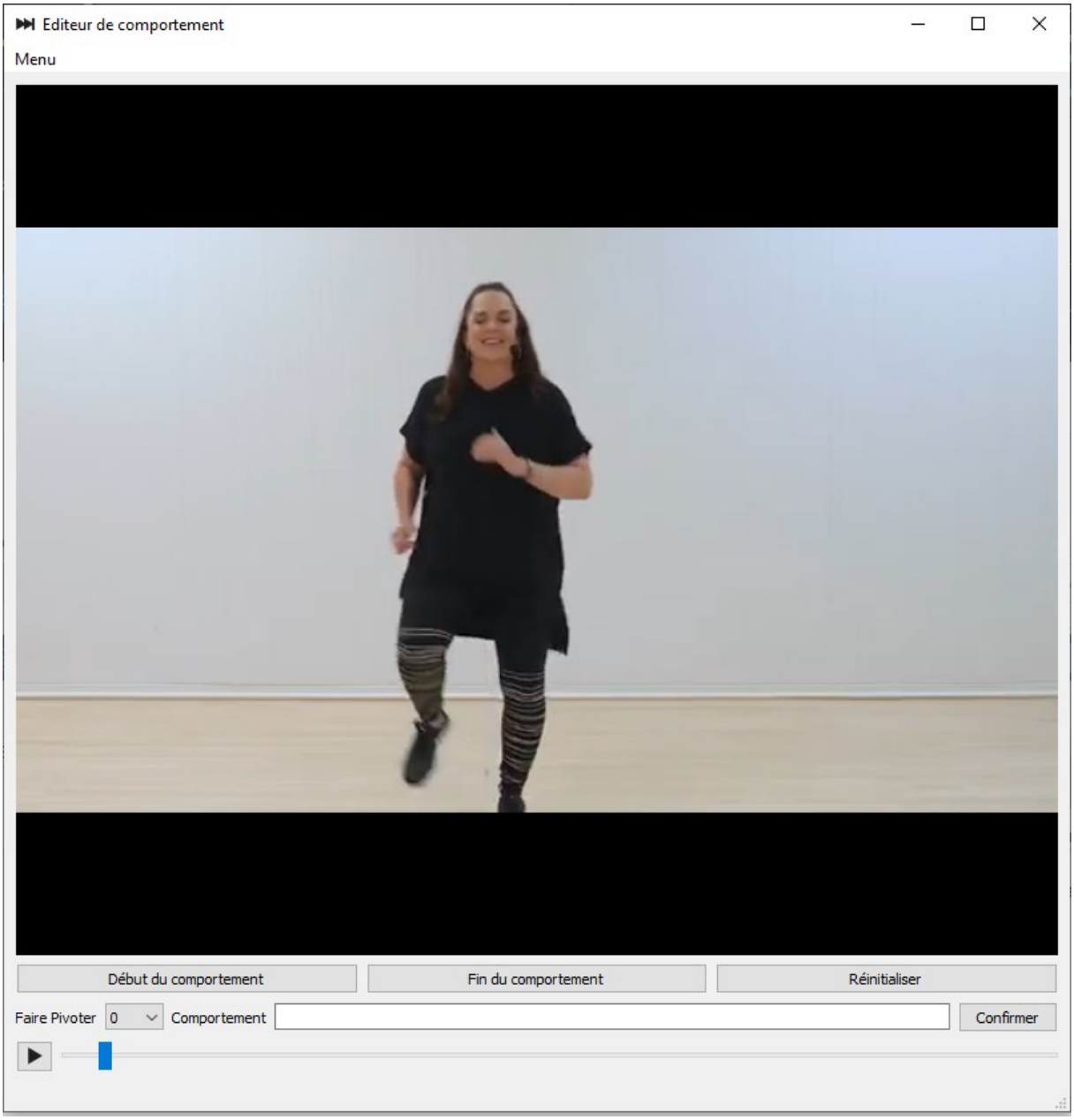

Figure 11. – Éditeur de comportement

À travers le système, l'usager est appelé à choisir le comportement qu'il veut observer. L'avatar reproduit les comportements choisis afin que l'usager puisse apprendre à reconnaitre le

comportement en question. Par la suite, l'usager est amené à interagir avec l'avatar en choisissant parmi les attitudes suggérées, celle qui convient le mieux dans la situation où un patient présente le comportement en question. Nous résumons l'utilisation du système sur la Figure 12.

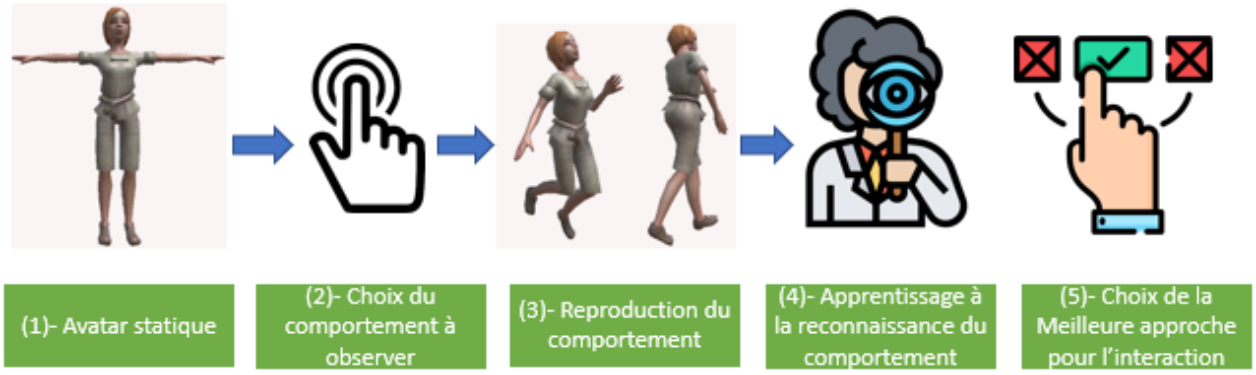

Figure 12. – Interaction avec le système de conseils

L'étape 4 de l'interaction avec le système de conseils (apprentissage à la reconnaissance du comportement) repose principalement sur l'utilisateur du système. Notre système permet d'améliorer l'apprentissage de comportements humains. Pour apprendre le comportement, l'utilisateur peut observer la reproduction du comportement de façon répétitive. Cette répétition permet de mémoriser les aspects du comportement et ainsi apprendre à le reconnaitre facilement.

La dernière étape (étape 5) qui consiste au choix de la meilleure approche, n'a pas fait l'objet de développent dans nos travaux. Cependant, nous établissons la logique à suivre pour la mettre en place. En effet, nous recommandons la réalisation d'une liste des approches appropriées pour chaque comportement reproduit. Ainsi, les approches de cette liste sont présentées individuellement a l'utilisateur. Le choix du l'approche adaptée entraînerait le retour de l'avatar à son état initial (étape 1 – Avatar statique).

6.5. Synthèse de la reproduction de comportements

Nous avons réalisé une reproduction de comportements en faisant un adressage des point clés extraits. Cet adressage a été réalisé par rapport au modèle standard d'un avatar humanoïde au sein de l'environnement Unity3D. Par la suite nous avons réalisé une coordination de mouvement

en calculant les rotations des points par rapport à un tableau d'hiérarchisation (Tableau 5). Nous avons mis en place un éditeur comportemental permettant d'étiqueter un certain nombre de comportement dans une vidéo afin de procéder à l'extraction de coordonnées des points clés pour chaque comportement étiqueté. Pour terminer, nous avons mis en place une interface permettant d'observer la reproduction de ces comportements à volonté. Afin d'interagir avec l'avatar dans le contexte d'apprentissage des approches par rapport aux comportements observés, nous avons proposé de présenter un ensemble d'approches. Notre proposition consiste à ramener l'avatar dans un état statique lorsqu'un utilisateur choisit la bonne approche à adopter par rapport à un comportement observé.

Chapitre 7 — Conclusion

À travers nos travaux, nous avons présenté une méthode pour animer un avatar virtuel en utilisant des coordonnées tridimensionnelles estimées à partir d'une vidéo. Dans le chapitre 2, nous avons fait un état de l'art des différentes approches suggérées pour l'estimation de la pose aussi bien bidimensionnelle que tridimensionnelle. Cet état de l'art nous a ensuite permis d'évaluer au chapitre 3, les approches les plus pertinentes aussi bien qualitativement que quantitativement. Concernant l'estimation 3D, nous avons évalué deux méthodes. Une première qui s'inscrit dans la catégorie des méthodes qui reposent sur l'inférence de coordonnées 3D à partir de coordonnées 2D et une seconde méthode qui appartient à la catégorie de méthodes qui estiment les coordonnées 3D par régression directe de l'image basée sur un modèle volumétrique de corps humain.

Au cours de notre évaluation de méthodes, nous avons constaté que les méthodes pour l'estimation de la pose 3D présentaient des résultats variables sur les images arbitraires en dépit des performances sur le jeu de donnes d'évaluation. La première méthode faisait un compromis sur la précision au profit de la rapidité de l'estimation. Nous avons également observé que la seconde méthode prônait la régression de la pose 3D sans perte d'informations sur l'image. Cependant, une limitation demeurait dans les cas d'images avec troncatures ou lorsque l'individu sur l'image ne se situait pas au centre de celle-ci. Grâce à notre évaluation des méthodes et la détermination des meilleures méthodes, nous avons pu formuler notre approche pour obtenir une estimation et l'extraction de la pose 3D. Ainsi, au chapitre 4, nous avons présenté notre approche qui se propose de soumettre chaque image d'une vidéo à une estimation 3D en 2 phases pour prédire la forme et la pose du corps humain. Au cours de la première phase, nous avons estimé les points clés en 2D en utilisant une architecture à deux branches qui utilise les champs d'affinité des parties du corps. Nous avons procédé au stockage des points clés estimés. Dans la deuxième phase, nous avons suggéré l'utilisation d'une méthode de reconstruction de maillage 3D pour régresser les points clés en 3D en utilisant les résultats de la première phase. Ce procédé se traduit par l'extraction et le stockage d'une séquence de points clés 3D (pour une vidéo

entière). Notre approche nous a permis d'améliorer l'estimation dans les vidéos à forte occlusion ainsi que dans le cadre d'images décentrées et présenter les résultats aussi bien pour la première phase que pour la seconde au chapitre 5.

Nous avons obtenu la reproduction des comportements exprimés dans des vidéos sur un avatar virtuel en exploitant les résultats de l'estimation de pose 3D. Cette reproduction a été discutée au chapitre 6 ainsi qu'une exploration de l'application de cette technique dans le domaine médical. Dans notre travail, nous nous sommes concentrés sur l'application pour la maladie d'Alzheimer. Nous avons mis en place un système reproduisant les comportements des patients atteints d'Alzheimer et prodiguant des conseils sur la façon de réagir face à un comportement particulier. Le but d'un tel système est d'éduquer le personnel médical dans son interaction avec les patients atteints de la maladie d'Alzheimer au cours de leurs épisodes.

À l'avenir, ce travail pourra être étendu dans de multiples directions. L'une serait d'intégrer la régression détaillée des points clés de la main. Cette estimation des poses des mains permettrait une reproduction plus complète du comportement de l'avatar. Une autre approche serait d'intégrer un encodeur temporel. Un encodeur temporel fournit une capture complète de la dynamique humaine qui aide à prédire le mouvement d'étouffement dans une vidéo, d'où une reproduction plus fluide du comportement. Une direction aussi fascinante serait d'étendre notre approche a l'estimation et l'extraction de pose de plusieurs individus au sein d'une image. Cette extension permettrait de réaliser de reproduction plus interactive de scènes en environnement virtuel et par la même occasion, enrichir notre système de conseils.

Références bibliographiques

1. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks.* Commun. ACM, 2017. **60**(6): p. 84–90.
2. Toshev, A. and C. Szegedy. *DeepPose: Human Pose Estimation via Deep Neural Networks*. in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
3. Carreira, J., et al. *Human Pose Estimation with Iterative Error Feedback*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
4. Szegedy, C., et al. *Going deeper with convolutions*. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.
5. Sun, X., et al. *Compositional Human Pose Regression*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
6. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
7. S, L.I., Z.Q. Liu, and A.B. Chan. *Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network*. in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014.
8. Gkioxari, G., et al., *R-CNNs for Pose Estimation and Action Detection.* arXiv e-prints, 2014: p. arXiv:1406.5212.
9. Xiaochuan, F., et al. *Combining local appearance and holistic view: Dual-Source Deep Neural Networks for human pose estimation*. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.
10. Wei, S., et al. *Convolutional Pose Machines*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
11. Newell, A., K. Yang, and J. Deng, *Stacked Hourglass Networks for Human Pose Estimation.* arXiv e-prints, 2016: p. arXiv:1603.06937.
12. Chu, X., et al. *Multi-context Attention for Human Pose Estimation*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
13. Yang, W., et al. *Learning Feature Pyramids for Human Pose Estimation*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
14. Chou, C., J. Chien, and H. Chen. *Self Adversarial Training for Human Pose Estimation*. in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2018.
15. Tang, W., P. Yu, and Y. Wu, *Deeply Learned Compositional Models for Human Pose Estimation*, in *15th European Conference on Computer Vision, ECCV 2018*, V. Ferrari, et al., Editors. 2018, Springer Verlag. p. 197-214.
16. Ren, S., et al., *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. **39**(6): p. 1137-1149.
17. He, K., et al. *Mask R-CNN*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.

18. Iqbal, U. and J. Gall. *Multi-person Pose Estimation with Local Joint-to-Person Associations*. in *Computer Vision – ECCV 2016 Workshops*. 2016. Cham: Springer International Publishing.
19. Fang, H., et al. *RMPE: Regional Multi-person Pose Estimation*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
20. Xiao, B., H. Wu, and Y. Wei. *Simple Baselines for Human Pose Estimation and Tracking*. in *Computer Vision – ECCV 2018*. 2018. Cham: Springer International Publishing.
21. Chen, Y., et al. *Cascaded Pyramid Network for Multi-person Pose Estimation*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
22. Moon, G., J.Y. Chang, and K.M. Lee. *PoseFix: Model-Agnostic General Human Pose Refinement Network*. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
23. Pishchulin, L., et al. *DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
24. Insafutdinov, E., et al., *DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model.* arXiv e-prints, 2016: p. arXiv:1605.03170.
25. Cao, Z., et al. *Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
26. Cao, Z., et al., *OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. **43**(1): p. 172-186.
27. Newell, A., Z. Huang, and J. Deng, *Associative Embedding: End-to-End Learning for Joint Detection and Grouping.* arXiv e-prints, 2016: p. arXiv:1611.05424.
28. Jin, S., et al., *Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation.* arXiv e-prints, 2020: p. arXiv:2007.11864.
29. Papandreou, G., et al. *PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model*. in *Computer Vision – ECCV 2018*. 2018. Cham: Springer International Publishing.
30. Yu, T., et al., *DoubleFusion: Real-Time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. **42**(10): p. 2523-2539.
31. Kadkhodamohammadi, A., et al. *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms*. in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017.
32. Zhi, T., et al. *TexMesh: Reconstructing Detailed Human Texture and Geometry from RGB-D Video*. in *Computer Vision – ECCV 2020*. 2020. Cham: Springer International Publishing.
33. Charles, R.Q., et al. *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
34. Jiang, H., J. Cai, and J. Zheng. *Skeleton-Aware 3D Human Shape Reconstruction From Point Clouds*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
35. Wang, K., et al. *Sequential 3D Human Pose and Shape Estimation From Point Clouds*. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

36. von Marcard, T., et al., *Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs.* Computer Graphics Forum, 2017. **36**(2): p. 349-360.
37. Zhang, Z., et al. *Fusing Wearable IMUs With Multi-View Images for Human Pose Estimation: A Geometric Approach*. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
38. Huang, F., et al. *DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image*. in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020.
39. Zhao, M., et al. *Through-Wall Human Mesh Recovery Using Radio Signals*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
40. Isogawa, M., et al. *Optical Non-Line-of-Sight Physics-Based 3D Human Pose Estimation*. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
41. Tome, D., et al. *xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
42. Saini, N., et al. *Markerless Outdoor Human Motion Capture Using Multiple Autonomous Micro Aerial Vehicles*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
43. Pavlakos, G., et al. *Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
44. Liang, J. and M. Lin. *Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
45. Kocabas, M., S. Karagoz, and E. Akbas. *Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry*. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
46. Remelli, E., et al. *Lightweight Multi-View 3D Pose Estimation Through Camera-Disentangled Representation*. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
47. Li, S., W. Zhang, and A.B. Chan. *Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation*. in *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015.
48. Pavlakos, G., X. Zhou, and K. Daniilidis. *Ordinal Depth Supervision for 3D Human Pose Estimation*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
49. Martinez, J., et al. *A Simple Yet Effective Baseline for 3d Human Pose Estimation*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
50. Tekin, B., et al. *Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
51. Zhou, K., et al. *HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
52. Jahangiri, E. and A.L. Yuille. *Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections*. in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017.

53. Sharma, S., et al. *Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
54. Li, C. and G.H. Lee. *Generating Multiple Hypotheses for 3D Human Pose Estimation With Mixture Density Network*. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
55. Loper, M., et al., *SMPL: a skinned multi-person linear model.* ACM Trans. Graph., 2015. **34**(6): p. Article 248.
56. Bogo, F., et al. *Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image*. in *Computer Vision – ECCV 2016*. 2016. Cham: Springer International Publishing.
57. Tan, V., I. Budvytis, and R. Cipolla, *Indirect deep structured learning for 3D human body shape and pose prediction*. 2017.
58. Pavlakos, G., et al. *Learning to Estimate 3D Human Pose and Shape from a Single Color Image*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
59. Omran, M., et al. *Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation*. in *2018 International Conference on 3D Vision (3DV)*. 2018.
60. Varol, G., et al. *BodyNet: Volumetric Inference of 3D Human Body Shapes*. in *Computer Vision – ECCV 2018*. 2018. Cham: Springer International Publishing.
61. Kanazawa, A., et al. *End-to-End Recovery of Human Shape and Pose*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
62. Mehta, D., et al. *Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision*. in *2017 International Conference on 3D Vision (3DV)*. 2017.
63. Nie, B.X., P. Wei, and S. Zhu. *Monocular 3D Human Pose Estimation by Predicting Depth on Joints*. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
64. Zhou, X., et al. *Deep Kinematic Pose Regression*. in *Computer Vision – ECCV 2016 Workshops*. 2016. Cham: Springer International Publishing.
65. Cheng, Y., et al. *Occlusion-Aware Networks for 3D Human Pose Estimation in Video*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
66. Xiang, D., H. Joo, and Y. Sheikh. *Monocular Total Capture: Posing Face, Body, and Hands in the Wild*. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
67. Joo, H., T. Simon, and Y. Sheikh. *Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
68. Wang, H., et al. *BLSM: A Bone-Level Skinned Model of the Human Mesh*. in *Computer Vision – ECCV 2020*. 2020. Cham: Springer International Publishing.
69. Zhou, X., et al. *Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
70. Wandt, B., H. Ackermann, and B. Rosenhahn, *3D Reconstruction of Human Motion from Monocular Image Sequences.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016. **38**(8): p. 1505-1516.
71. Rhodin, H., et al. *General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues*. in *Computer Vision – ECCV 2016*. 2016. Cham: Springer International Publishing.

72. Huang, Y., et al. *Towards Accurate Marker-Less Human Shape and Pose Estimation over Time*. in *2017 International Conference on 3D Vision (3DV)*. 2017.
73. Peng, X.B., et al., *SFV: reinforcement learning of physical skills from videos.* ACM Trans. Graph., 2018. **37**(6): p. Article 178.
74. Ionescu, C., et al., *Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014. **36**(7): p. 1325-1339.
75. Lin, T.-Y., et al. *Microsoft COCO: Common Objects in Context*. in *Computer Vision – ECCV 2014*. 2014. Cham: Springer International Publishing.
76. Nair, V. and G.E. Hinton, *Rectified linear units improve restricted boltzmann machines*, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 2010, Omnipress: Haifa, Israel. p. 807–814.
77. Simonyan, K. and A. Zisserman *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. arXiv:1409.1556.
78. Russakovsky, O., et al., *ImageNet Large Scale Visual Recognition Challenge.* International Journal of Computer Vision, 2015. **115**(3): p. 211-252.
79. Sandler, M., et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.
80. Howard, A.G., et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv:1704.04861.
81. Abadi, M., et al., *TensorFlow: a system for large-scale machine learning*, in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. 2016, USENIX Association: Savannah, GA, USA. p. 265–283.
82. Koks, D., *Explorations in Mathematical Physics: The Concepts Behind an Elegant Language*. 2006: Springer Science+Business Media,LLC. Chapter 4, page 147.

Annexes

Annexe A

Nous avons soumis un article court accepté à la 17e conférence internationale sur les systèmes de tutorat intelligent intitulé « INTELLIGENT TUTORING SYSTEMS IN AN ONLINE WORLD » tenue à Athènes, Grèce en Juin 2021 (Springer Verlag, Lecture notes in Computer Science).

La contribution de chaque auteur se répartit comme suit :

Kodjine Dare : Étude des travaux dans le domaine, élaboration de la méthode, mise en place des expérimentations et rédaction de l'article.

Hamdi Ben Abdessalem : orientation sur le fonctionnement des avatars en environnement virtuel, coordination des travaux et relecture de l'article avant soumission

Claude Frasson : formulation du sujet de recherche, relecture de l'article et financement des travaux.

Extraction of 3D Pose in Video for Building Virtual Learning Avatars

Kodjine Dare, Hamdi Ben Abdessalem and Claude Frasson

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, Montréal, Canada H3C 3J7
{kodjine.dare,hamdi.ben.abdessalem}@umontreal.ca,frasson @iro.umontreal.ca

Abstract. From an image of a person, we can easily guess the 3D coordinates of the body parts. This is because we have acquired a 3D mental model from observing humans and interacting with them. This capacity easily achievable for humans is not systematic when it comes to computers. In this paper, we describe an approach that aims at estimating poses from video with the objective of reproducing the observed movements by a virtual avatar. We propose the fragmentation of submitted videos into series of RGB frames to process individually. We aim two main objectives in our work. First, we achieve the extraction of initial 2D joints coordinates using a method that predicts joint locations by part affinities (PAFs). Then we infer 3D joints coordinates based on a human full 3D mesh reconstruction approach supplemented by the previously estimated 2D coordinates. Secondly, we explore the reconstruction of a virtual avatar using the extracted 3D coordinates with the prospect to transfer human movements towards the animated avatar. This would allow to extract the behavioral dynamics of a human, allowing to detect some health problems, for instance in Alzheimer. Our approach consists of multiple subsequent stages that show better results in the estimation and extraction than similar solution due to this supplement of 2D coordinates. With the final extracted coordinates, we apply a transfer of the positions (per frame) to the skeleton of a virtual avatar in order to reproduce the movements extracted from the video.

Keywords: Machine Behavior, Behavior detection, Visualization, 3D Pose Estimation, Virtual Avatar.

Introduction

Considering a video that displays a person performing a task or even just a movement (walking, running, dancing), human can easily identify the different body parts location and orientation of the person also known as pose in the video. The analysis of this simple human process introduces the interrogation regarding the capacity of computers to automatically detect human body pose.

It is in this same vein that the purpose of this article revolves around two focal points. Firstly, we articulate our work around this task known as Human pose estimation that aims to automatically locate the human body parts from images or videos in order to extract information such as human poses. This human pose estimation is also known as **keypoints detection**. The extraction of these poses provides a collection of data that will be relevant for the next stage. Then we address the second objective of our work that consists in using the extracted body part location for the reproduction of the behavior with an animated avatar. We define the animation of an avatar as a sequence of poses extracted in a video. This objective aims to allow to transfer human movements towards an animated avatar to highlight the behavioral dynamics of the human from the keypoints that have been extracted.

The achievement of these two objectives can lead to multiple implications in both applied and theoretical research. In the context of this research, we focus on the application for Alzheimer behavior. Alzheimer is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills and affects behavior. Alzheimer patients can have sometimes specific behavior (walking, equilibrium,…) which could be observed by video camera at different times of the day, extracted, and rebuilt in a virtual avatar. This avatar would serve as a training model to educate the medical staff to recognize an episode of Alzheimer patients.

Throughout this paper, we put special emphasis on the extraction of human pose estimation. We use in the rest of this paper the term pose and keypoints interchangeably. The remaining of this paper is organized as follows. In section 2 we present related works. In section 3 we present our approach and discuss the results in section 4.

Related Work

To address human pose estimation, several approaches have been proposed, we focus our discussion on relevant 2D and 3D human pose estimation.

Human 2D pose estimation: There are single person pose estimation methods and multi-person pose estimation methods. The single person methods are divided between the heatmap approach that chooses the locations with the highest heat values as the keypoints [1] and the direct regression approach which utilize the output feature maps to regress keypoints directly [2]. Multi-person methods host two categories: top-down approaches which consist of applying a person detector and then running a pose estimation algorithm per every detected person ([3], [4]) and bottom-up approaches which first step is to locate all the keypoints in an image and the second step is to group them according to the person they belong to [6].

3D pose estimation: Agarwal and Triggs [7] rely on silhouette feature while Zhou et al. [8] proceed by manual interaction from user. With deep learning, direct regression approaches integrate the SMPL [9] to train models to directly infer the SMPL parameters. These methods infer the 3D pose and shape based on: RGB image as suggested by Kanazawa et al. [10], RGB image and 2D keypoints [11], keypoints and silhouettes [12], or keypoints and body part segmentations [13]. Some methods extend their work to estimate 3D pose from **video**. Among these methods, the vast majority rely on elaborate environment which capture sequences on multiple angles. Due to the perspective of our work, we focus on the approaches that deal with video captured by regular cameras. Some methods obtain accurate shapes and textures of clothing by pre-capturing the actors and making use of silhouettes [14]. While these approaches obtain satisfying shape, reliance on the pre-scan and silhouettes restricts these approaches to videos obtained in an interactive and controlled environment. Therefore, we propose an approach that rely on RGB image supplemented by 2D keypoints.

Proposed Approach

The approach that we propose estimates 3D keypoints through a 2-stages estimation of frames of the input video and a transfer of results towards reconstruction. We propose the fragmentation of the video into series of RGB frames (see Fig. 1).

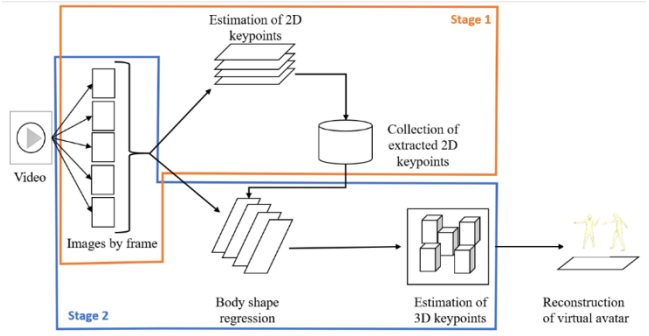

Fig. 1. Architecture of the proposed approach

First stage: It insures the 2D pose estimation using the Realtime Multi-Person 2D Pose Estimation [6] to estimate the human body 2D keypoints in the submitted frames.

This method presents a bottom-up approach for estimation of multi-person pose in RGB image and produces, as output, the 2D locations of anatomical keypoints for each person in the image, without using person detector. The model defines a network architecture that iteratively predicts affinity fields that encode part-to-part association and detection confidence maps. The network is split into two main sections: a first one that predicts the confidence maps, and the second predicts the affinity fields. Each section is an iterative prediction architecture, that finetunes the predictions throughout multiple stages, with intermediate supervision at each stage.

Once the estimation completed, we proceed to the extraction of the estimated keypoints. The result of such operation is a collection of keypoints per frame.

Second stage: We use End-to-end Recovery of Human Shape and Pose [10] to iterate the feature of every frame to predict the human body. The method infers a full 3D mesh of a human body directly from an RGB image of a human. With that approach, the 3D mesh of a human body is encoded using SMPL which generates human bodies into shape with regards to the variation in height, weight and body proportion, and the deformation of the surface due to the movements.

While this 3D estimation method presents a great approach to infer the human body shape from an RGB image, this method is introduced to work on input images of a particular scale and quality. To address that limitation, we strengthen the prediction by using the output of the first stage. This helps to estimate the final 3D keypoints that fits as much as possible the size of individuals in every frame. This process results in the extraction and storage of 3D keypoints that will be further used in a virtual environment to reconstruct an avatar which keypoints would correspond to the extracted keypoints. The overall idea behind the presented pipeline is to take a video as input and produce a virtual avatar that replicates the movement observed in the submitted video.

Results and Discussion

2D estimation

We conducted the experiment with the objective of estimating accurately human 2D pose from single RGB image regardless of the complexity of pose described by the image. In order to determine how adequate the method was for the purpose of our work, we have analyzed the results obtained and compared with the performance of similar methods. The methods compared were AlphaPose [4], PersonLab [15], Mask R-CNN [3], Deepcut [5], Stacked Hourglass Network [1]. While DeepCut achieved average performance, the qualitative results were by far the poorest. Stacked Hourglass Network and Mask R-CNN presents improved results on the qualitative and quantitative level compared to the previous one but did not deliver optimal performance. AlphaPose, PersonLab and our adopted approach gave the most consistent results and high amelioration.

In Fig. 2 we can observe the estimation difference between some of the methods we have tested and evaluated. The first column represents the estimation of the method we opted for, the second column the estimation of AlphaPose, the third one is the performance of PersonLab and the last represent the performance on Mask R-CNN.

The performance of AlphaPose and the adopted approach are not so far from each other, but we observe a rapid decay of the estimation when there is an occlusion of the human pixels with the background pixels.

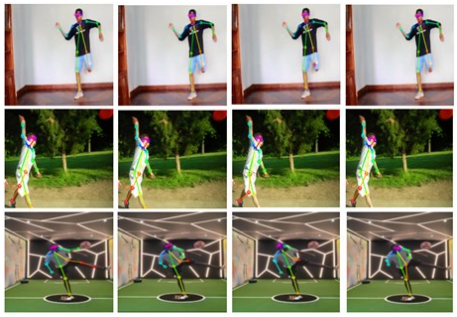

Fig. 2. Estimation across set of methods

3D estimation

The proposed approach was evaluated on different format of images. The objective that drove the assessment of the method was to find out the scope of the improvement of the results in comparison to the initial approach. Hence, we first measured the achievement of the model as suggested by Kanazawa et al. [10] and the model based of 2D keypoints the same set of images to see if our approach gives better results. We have not noticed a significant variation in the mesh inference, hence the 3D keypoints coordinates. This insignificant variation could be explained by the fact that regardless of the presence of 2Dkeypoint given that the initial approach performed best on images of with such scale, providing a base 2D keypoints to help determine the location of individual does not change much the outcome.

We have also evaluated the results of our approach on images of different scale to compare the mesh reconstruction. We represent the inference without the context provided by prior 2D estimation by a mesh in magenta, and in blue the inference with prior 2D keypoints.

In Fig. 3, we can observe the marge of variation performance for slight improvement to obvious differences. The first row shows how the most difference, without the initial context provided by the 2D keypoints, the model is not able to infer the 3D keypoints. The last row shows how well the model performs when it comes to reconstruct a limb that is out of vision range.

Fig. 3 Performance on images of different scale and bounding box

This finalized approach will be used to reproduce movement by avatar. These avatars will be leveraged in the context of education systems. The idea behind these systems will be to allow learners to interact with avatars that display some behaviors and visualize the response of their interactions.

Conclusion

In this paper, we proposed an approach to animate a virtual avatar based on 3D keypoints estimated from a video. Our proposed approach divides the video into series of RGB images, and for each image we suggested to perform a 2-stages estimation. During the first stage, we have estimated 2D keypoints using a 2D pose estimation and we have proceeded to the extraction of the estimated keypoints. In the second stage, we have used a 3D mesh reconstruction method to infer 3D keypoints using the output of the first stage. This process results in the extraction and storage of a sequence of 3D keypoints. We use this sequence to reproduce the movement from the video on a virtual avatar. With the proposed solution, we were successfully able to reproduce the video behavior on an avatar in a virtual environment. This approach could have multiple applications, but in the context of this research, we focus on the application for Alzheimer's disease. In fact, such solution is devoted to help in the creation of a system by reproducing Alzheimer's disease patient's behavior on a virtual avatar to educate medical staff in the interaction with them.

Acknowledgment. We acknowledge NSERC-CRD (National Science and Engineering Research Council Cooperative Research Development), Prompt, and BMU (Beam Me Up) for funding this work.

References

A. Newell, K. Yang, J. Deng: Stacked Hourglass Networks for Human Pose Estimation. In: Leibe B., Matas J., Sebe N., Welling M. (eds) ECCV 2016, LNCS, vol 9912, pp. 483-499. Springer, Cham (2016).

A. Toshev, C. Szegedy: DeepPose: Human Pose Estimation via Deep Neural Networks. In: Proceedings of the IEEE Conference on CVPR, pp. 1653-1660. IEEE Computer Society, Las Vegas-USA (2014).

K. He, G. Gkioxari, P. Dollár, R. Girshick: Mask R-CNN. In: ICCV, pp. 2980-2988. IEEE Computer Society, Venice-Italy (2017).

H. Fang, S. Xie, Y.-W. Tai, C. Lu: RMPE: regional multi-person pose estimation. In: ICCV, pp. 2353–2362. Venive-Italy (2017).

L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. An-driluka, P. V. Gehler, B. Schiele: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: IEEE conference on CVPR, pp. 4929–4937. USA (2016).

Z. Cao, T. Simon, S. Wei, Y. Sheikh: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on CVPR, pp. 7291-7299. Hawaii (2017).
A. Agarwal, B. Triggs: Recovering 3d human pose from monocular images. TPAMI 28(1), 44–58(2006).
S. Zhou, H. Fu, L. Liu, D. Cohen-Or, X. Han: Parametric reshaping of human bodies in images. In: SIGGRAPH '10, pp. 1-10. Association for Computing Machinery, USA (2010).
M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black: SMPL: A skinned multi-person linear model. ACM Trans. Graphics 34(6), 1-16 (2015).
A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on CVPR, pp. 7122-7131. IEEE Computer Society, Salt Lake City-USA (2018).
H.-Y. Tung, H.-W. Tung, E. Yumer, K. Fragkiadaki: Self-supervised learning of motion capture. In: Proceedings of the 31st International Conference on NIPS, pp. 5242–5252. Curran Associates Inc, Long Beach-USA (2017).
G. Pavlakos, L. Zhu, X. Zhou, K. Daniilidis: Learning to estimate 3D human pose and shape from a single-color image. In: Proceedings of the IEEE Conference on CVPR, pp. 459-468. IEEE Computer Society, Salt Lake City-USA (2018).
M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, B. Schiele: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conference on 3DV, pp. 484-494. IEEE Computer Society, Verona-Italy (2018).
T. Alldieck, M. Magnor, W. Xu, C. Theobalt, G. Pons-Moll: Video based reconstruction of 3d people model. In: Proceedings of the IEEE Conference on CVPR, pp. 8387-8397. IEEE Computer Society, Salt Lake City-USA (2018).
G. Papandreou, T. Zhu, Tyler, L-C. Chen, S. Gidaris, J. Tompson, K. Murphy: PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) ECCV 2018, LNCS, vol 11218. Springer, Cham (2018).