Titre: Title:	Génération d'animations par capture de mouvements avec multiples caméras 3D
Auteur: Author:	David Ménard
Date:	2014
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ménard, D. (2014). Génération d'animations par capture de mouvements avec multiples caméras 3D [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/1647/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/1647/
Directeurs de recherche: Advisors:	Benoît Ozell
Programme: Program:	Génie informatique

UNIVERSITÉ DE MONTRÉAL

GÉNÉRATION D'ANIMATIONS PAR CAPTURE DE MOUVEMENTS AVEC MULTIPLES CAMÉRAS 3D

DAVID MÉNARD
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)
DÉCEMBRE 2014

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

GÉNÉRATION D'ANIMATIONS PAR CAPTURE DE MOUVEMENTS AVEC MULTIPLES CAMÉRAS 3D

présenté par : MÉNARD David
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. BOYER François-Raymond, Ph. D., président
M. OZELL Benoît, Ph. D., membre et directeur de recherche
M. HURTUT Thomas, Ph. D., membre

DÉDICACE

À mon professeur, Benoît Ozell,
et à mon superviseur, André Foisy,
pour leur support et motivation ...

REMERCIEMENTS

J'aimerais remercier Benoît Ozell pour sa supervision et direction de la recherche, André Foisy pour sa supervision et Autodesk Inc pour son soutien matériel et financier.

RÉSUMÉ

Le présent mémoire traite du sujet de la capture de mouvement par caméra de profondeur. Plus précisément, la recherche émet l'hypothèse que la capture de mouvement produite par une simple caméra 3D peut être améliorée en intégrant multiples caméras et un système de cinématique inverse dans une solution de capture complète.

La solution produite intègre plusieurs composantes indépendantes, qui intègrent le modèle humain de la bibliothèque HumanIK d'Autodesk dans un système de capture utilisant multiples caméras Kinect.

Une architecture client-serveur est utilisée afin de combiner les données de multiples caméras 3D, en transformant le système de coordonnées de chacune d'elle dans une référence centrale. Nous établissons qu'un maximum de 3 caméras peuvent être utilisées simultanément, et procédons ensuite à combiner les squelettes en effectuant une moyenne intelligente sur les donnes accumulées.

La bibliothèque de cinématique inverse est ensuite intégrée au système. Plusieurs possibilités d'intégration sont explorées et mesurées. Finalement, l'application de la cinématique inverse après la capture du squelette à chaque trame est retenue comme étant la méthode la plus précise. De plus, toutes les fonctions dérivées de la cinématique inverse, tel que les limites de flexibilité du corps humain et la correction de poses sont retenues dans la solution, car elles améliorent significativement la qualité des séquences.

Finalement, de l'information supplémentaire pour compléter la cinématique inverse est recherchée dans les images de profondeur. De l'information sur l'état des mains du sujet et sur l'orientation de la tête est extraite et intégrée dans le modèle humain de HumanIK.

Avec une analyse quantitative et qualitative de chaque composante indépendante et combinée, la recherche montre que la séquence de mouvement finale produite montre une amélioration significative par rapport à une simple capture par caméra 3D.

Malgré les résultats montrant une amélioration significative des captures produites, celles-ci demeurent de qualité insuffisante afin d'être utilisées dans un contexte pratique. Par contre, l'intégration de la recherche dans le logiciel *Motion Builder* d'Autodesk a été faite et montre un potentiel et une grande attente par rapport à la capture de mouvement par caméra 3D.

ABSTRACT

This research looks to enhance the quality of motion capture generated by a single depth camera. It suggests that by using multiple Kinect cameras as input and combining a human model of inverse kinematics, a motion sequence can be significantly improved.

The first step into producing the proposed solution aims to reconcile the data from multiple depth cameras. These cameras are active sensors and interfere with each other, so that their use is limited. We establish that a higher limit of 3 cameras can be used simultaneously in our conditions and proceed to combine their data into a single output skeleton. By converting each cameras transformation space into a centralized reference, we can perform a smart average on each skeleton and use the resulting data as a basis on which to apply the inverse kinematics integration.

Using the HumanIK library by Autodesk, the research then explores multiples ways of integrating an inverse kinematics solution into the motion capture pipeline. We determine that applying the kinematic equations after each pose has been calculated at each frame produces the best results. Furthermore, each function offered by HumanIK, derived from inverted kinematics applied to a human model, is analyzed. We find that all the functions, such as human flexibility models and pose correction, have benefits on the resulting sequence of movements.

Finally, we look for additional data in the depth images to enhance the information fed to our IK model. By extracting hand states and head position and rotation, we can use HumanIK to calculate and guess at these values, significantly enhancing the quality of the produced MoCap sequences.

We finally perform both quantitative and qualitative measurements. To do so, we compare our results with a commercial *Flock Of Birds* system and use a group of individuals to perform a blind test on resulting animations. We find that we can significantly enhance both the accuracy and quality of animation produced compared to that produced with a single depth camera.

Despite the results being significantly better, the animations created are still not to a quality standard that can be used in a production pipeline. Despite this, the solution has been integrated in Autodesk *Motion Builder* software showing big promise, potential and expectancy towards depth camera based motion capture.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

HumanIK	*Human Inverted Kinematics*
IK	*Inverted Kinematics*, cinématique inverse
Mocap	*Motion Capture*, capture de mouvement

CHAPITRE 1 INTRODUCTION

1.1 Introduction

La présente recherche a été déclenchée par la démocratisation des caméras 3D dans le domaine du divertissement, particulièrement dans le domaine du jeu vidéo. L'utilisation des caméras de profondeur est encore très limitée et leur valeur peut être utilisée dans plusieurs domaines, dont la capture de mouvements.

1.2 Contexte

Ce travail de recherche est encadré par l'École Polytechnique de Montréal, supervisé par Benoît Ozell, et en collaboration avec la compagnie Autodesk Inc. Ce contexte particulier a permis de bien faire évoluer plusieurs aspects du projet.

Tout d'abord, l'encadrement fourni à l'École Polytechnique a permis d'assurer une approche méthodique, centrée sur la recherche et l'avancement du domaine scientifique. Cette approche assure la réponse aux questions avancées lors de la formulation du problème de recherche et définit des métriques mesurables et comparables.

L'encadrement de la compagnie Autodesk Inc. a permis de maintenir une approche pratique à la recherche et d'avoir un soutien matériel et financier afin de permettre sa réalisation. Cette collaboration a facilité l'accès aux informations pertinentes, au matériel requis et à plusieurs personnes ressources. Ces dernières ont aidé à guider la recherche afin que celle-ci soit adaptée à un environnement de production et ont pu évaluer la qualité des résultats obtenus.

1.3 Convergence du sujet

1.3.1 Questions théoriques

Visant à améliorer les techniques de capture de mouvement à l'aide de caméras 3D, la recherche soulève plusieurs questions théoriques pertinentes. Les caméras 3D étant très récentes et l'incertitude de leurs mesures étant de l'ordre de plusieurs centimètres, une capture de mouvement effectué avec celles-ci sera également de basse qualité, comparée à des systèmes de capture de mouvements commerciaux. Les questions qui sont soulevées par cette recherche dérivent toutes de l'objectif qui est d'améliorer les résultats obtenus directement avec une caméra 3D et ses logiciels d'opération.

La première question est soulevée lorsque l'on intègre plus d'une caméra 3D durant le recueil de données pour la capture de mouvement. En effet, il est intuitif de croire que l'addition de caméras 3D pour aider à la capture de mouvement ne peut qu'améliorer les résultats de l'animation résultante. Par contre, ces caméras émettent une lumière infrarouge, toutes de la même longueur d'onde, il y a donc une possibilité d'interférence entre les caméras. Il est donc intéressant de quantifier l'interférence et observer les répercussions de celles-ci sur la détection du squelette du sujet. Ceci permettra alors de mesurer l'effet de multiples caméras 3D sur la qualité et la précision de la capture de mouvement effectué. Les réponses à ces questions feront alors partie des résultats de la recherche et permettront de quantifier les conséquences d'ajouter des caméras 3D sur l'animation finale capturée.

La deuxième question porte sur l'effet d'intégrer la cinématique inverse dans le modèle du squelette fourni par les caméras 3D sur le résultat de la capture. Le squelette extrait des images de la caméra 3D est calculé avec des approches d'analyse d'images et ne comporte aucune restriction liée au corps humain. La bibliothèque HumanIK d'Autodesk implémente la cinématique inverse et intègre les contraintes de celle-ci adaptées au corps humain. Ainsi, nous allons expérimenter sur les conséquences d'ajouter un modèle du corps humain basé sur la cinématique inverse aux captures sortant des caméras 3D. Par leur quantification et en mesurant leurs effets sur la précision et la qualité visuelle, nous pourrons déterminer si la cinématique inverse a un effet bénéfique sur la capture de mouvement assistée par caméras de profondeur.

1.3.2 Applications pratiques

Le projet de recherche étant encadré par la compagnie Autodesk Inc., ses applications pratiques prennent une place importante afin d'avoir le potentiel d'être intégré dans un ou plusieurs de leurs produits commerciaux. Les caméras 3D sont maintenant accessibles au public et sont peu dispendieuses comparé à un système de capture de mouvement commercial. Ceci représenterait un avantage significatif si la capture de mouvement produite avec ces caméras était de qualité acceptable. Malheureusement, les séquences de mouvements capturées par les caméras 3D et leur logiciel de traitement sont imprécises et montrent beaucoup d'imperfections. Ces problèmes font que leurs utilisations dans des applications commerciales sont impossibles. En effet, la qualité des animations produite n'est pas assez bonne pour justifier le coût de ses retouches par un animateur professionnel. Ceci pose un problème majeur pour les studios de développement de jeux vidéo qui ont un budget serré qui ne peut pas se procurer des solutions de capture de mouvements. En cherchant à améliorer de façon significative les résultats de la capture de mouvement avec des caméras 3D, cette recherche peut être la base d'une implémentation de capture de mouvement simple, rapide et peu dispendieux dans les produits d'Autodesk.

1.3.3 Faisabilité et limites attendues

Les caméras 3D étant très récentes, elles présentent plusieurs limites pratiques. Tout d'abord, les pilotes logiciels associés à ces dispositifs sont incapables de supporter plus d'une caméra 3D par contrôleur USB. De plus, l'implémentation actuelle de leur interface de programmation d'application ne permet pas d'initialiser plus d'une caméra 3D par processus [1]. Ces limites peuvent être contournées en utilisant une architecture client-serveur pour rassembler les données de chaque dispositif. La deuxième frontière à franchir est celle imposée par la fiabilité de l'interprétation des données de la caméra 3D. La reconnaissance de squelette des bibliothètes utilisé est très sensible à l'occlusion des membres du sujet et à l'interférence entre les caméras 3D. Une fois ces deux problèmes surmontés, le cœur de la recherche, qui est d'améliorer les résultats des animations résultantes de la capture de mouvements à l'aide de caméras 3D, peut prendre place.

1.4 Structure et descriptions

1.4.1 Méthodologie et résultats complémentaires

Le chapitre 3 présente la méthodologie utilisée permettant de mener l'expérience qui saura confirmer ou rejeter l'hypothèse émise. Les différentes étapes nécessaires pour la combinaison de l'information de multiples caméras de profondeur sont suggérées. Ensuite, des possibilités sur l'intégration de la cinématique inverse dans le pipeline de capture de mouvements sont suggérées.

À partir des explorations faites, un plan est créé et une méthode de validation des résultats est mise en place. La méthodologie établie dans cette section servira de plan durant l'expérimentation et la validation des résultats.

1.4.2 Articles

Notre article *Motion Capture Using Multiple Depth-Cameras Integtrated With Inverted Kinematics* présenté au chapitre 5 élabore la méthode utilisée pour implémenter la technique

---

[1] Cette limitation est maintenant obsolète. Plusieurs caméras Kinects peuvent maintenant être initialisées par processus, mais la limitation d'une caméra par contrôleur USB demeure présente

proposée. L'article soulève les problèmes rencontrés et les méthodes de résolution. Les différentes composantes de la technique sont ensuite exposées et les méthodes d'analyse sont détaillées. Finalement, les résultats obtenus sont présentés et analysés.

1.4.3 Discussion

Le chapitre 6 reprend les résultats présentés dans l'article et en fait une analyse plus profonde. D'abord, le sujet de la précision de la technique de capture de mouvements suggérée est abordé. Avec les résultats en main, nous pourrons tirer les conclusions appropriées quant à l'amélioration quantitative des séquences de mouvements produits.

Ensuite, nous observerons les résultats qualitatifs de chaque composante du système, afin d'en déterminer les limites et imperfections visuelles, ainsi que les améliorations apparentes qui sont amenées par rapport à la capture par simple caméra de profondeur.

Finalement, une analyse pratique de la technique sera faite, afin de déterminer si la technique proposée génère des animations de qualité suffisante pour être utilisée dans un contexte de production.

1.4.4 Avancements et travaux futurs

Enfin, le chapitre 7 fait une analyse des limites rencontrées durant la recherche. Cette analyse comprendra une vision des recherches concurrentes à celle-ci et permet d'établir quelles améliorations peuvent être amenées à la technique suggérée et quelles recherches supplémentaires restent à faire.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Introduction

Les caméras 3D ont déjà été le sujet de plusieurs recherches au cours des dernières années. Elles sont maintenant disponibles au grand public à un prix abordable. De plus, avec l'avènement de la caméra Kinect de Microsoft, leur adoption est de plus en plus populaire pour des fins de divertissements. En plus de recherche directe sur la technologie derrière ce type de caméras, de nombreuses recherches ont été faites sur leur utilisation afin de reconnaître un être humain et d'en extraire un squelette interactif.

D'un autre côté, les techniques de capture de mouvements continuent d'évoluer. Les solutions les plus robustes utilisent différents marqueurs et senseurs afin de déterminer la position des points-clés sur le sujet capturé. Certaines techniques de capture de mouvements sans marqueurs ont également fait surface, mais leurs précision et utilité pratique ne sont pas comparables aux techniques avec marqueurs.

L'utilisation de caméras 3D dans un contexte de capture de mouvement a été brièvement explorée. Le manque de précision de ces caméras ne justifie pas leur utilisation dans un milieu pratique.

D'autres recherches ont été faites pour combiner les données de plusieurs caméras 3D afin de raffiner le positionnement des squelettes extrait des images de profondeur. Les limitations des pilotes des caméras 3D rendent cette tâche difficile. De plus, l'interférence entre les différentes caméras elles-mêmes cause certaines limites par rapport à leur utilisation simultanée. Aussi, l'utilisation de plusieurs de ces caméras demande un travail de fusion de données afin d'utiliser efficacement les données en trois dimensions de chacune des caméras.

Finalement, aucun travail n'a été fait sur la combinaison de multiples caméras 3D avec un système de cinématique inverse, tel que présenté dans cette recherche. Par contre, l'intégration de modèles physiques et humains dans la capture de mouvement par caméras 3D a été recherchée. Ces techniques, similaires à l'approche présentée dans cette recherche, ont le potentiel de très bien synergiser avec la technique présentée.

2.2 Caméras 3D

Les caméras de profondeur produisent des images où la valeur de chaque pixel représente sa distance à la caméra (Figure 2.1). Ces images de profondeur ont l'avantage de montrer de

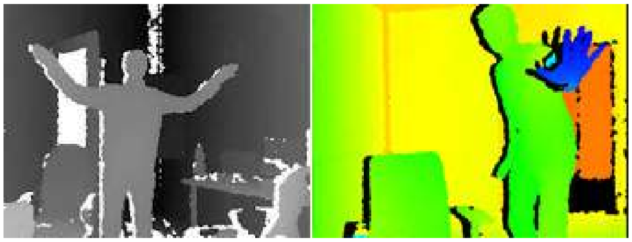

Figure 2.1 Image de profondeur générée par la caméra 3D Kinect de Microsoft

très fortes variations entre l'avant-plan et l'arrière-plan d'une scène.

Pour cette raison, les images de profondeur sont beaucoup mieux adaptées pour reconnaître des sujets que des images RGB produites par des caméras standards. Les contrastes élevés des contours des sujets permettent non seulement une distinction plus évidente, mais demandent également beaucoup moins de calculs afin d'isoler les figures intéressantes. Une fois l'image de profondeur obtenue, un *Middleware* est appelé à y extraire l'information désirée. Dans la plupart des cas, un squelette humain est obtenu.

Les caméras 3D telles que la *Kinect de Microsoft* et la *Xtion Pro de Asus* émettent un laser infrarouge qui rebondit sur les objets de la scène. La lumière émise est une lumière dite "structurée" ; un motif de lumière est émis et la distance est déduire selon la déformation de ce motif. Les bibliothèques OpenNI et KinectSDK retournent le résultat dans une image de profondeur en niveaux de gris ou une image RGB.

Il existe deux grands joueurs dans le monde des caméras 3D ; *Microsoft Kinect SDK* Microsoft Corp (2012) et *OpenNI* (Open Natural Interface) PrimeSense (2012). Les deux fournissent des pilotes pour la caméra *Kinect* de la console de jeu *XBox360* et *Kinect* pour Windows. De plus, OpenNI supporte également d'autres périphériques tel que la caméra 3D *Xtion Pro* de Asus, qui fournit 60 images par secondes au lieu des 30 de la caméra *Kinect*.

Les deux *SDK* fournissent également leur traitement afin d'extraire le squelette d'un ou plusieurs sujets se trouvant dans la scène observée. Les deux utilisent des méthodes similaires, mais pour l'instant, seul *Kinect SDK* donne la position des paires poignets/mains et chevilles/pied ; *OpenNNI* ne fournit que la position des mains et pieds. Des plans sont faits pour supporter ces jointures dans OpenNI sous peu. Par contre, l'architecture modulaire de OpenNI lui donne un avantage distinct. Étant *Open Sourced* OpenNI permet facilement de créer et ajouter des modules indépendants au *Middleware* pour des tâches spécifiques. Par exemple, il existe un module pour faciliter le suivi des mains et la reconnaissance de gestes.

2.2.1 Reconnaissance de squelettes

Les détails d'implémentation de l'algorithme de détection de squelettes de la première version du *Kinect SDK* ont été publiés en 2011 Shotton et al. (2011). Leur méthode propose de mapper le problème complexe d'estimation de pose à un problème plus simple de classification de pixels clés. Ainsi, en utilisant une très grande banque d'images comme échantillons d'apprentissage (plus de 300000 images), leur méthode permet de détecter les membres individuels d'une pose complète de façon très efficace et précise.

Les techniques plus récentes du *Kinect SDK* ne sont pas publiées et demeurent secrètes. Il en est de même pour l'approche utilisée par *PrimeSense* dans *OpenNI.* Les méthodes de détection de squelette humain des bibliothèques de Microsoft et OpenNI utilisent des méthodes d'analyses d'images découvertes par l'analyse d'images RGB.

Par contre, leurs méthodes n'utilisent pas d'heuristiques et d'analogues d'humains. En effet, un modèle du corps humain et de ses limitations n'a jamais été intégré dans une solution de détection de pose, autant dans le domaine de traitement d'images RGB classiques que d'images de profondeur.

2.2.2 Multiples caméras

Dès l'apparition de la caméra *Kinect* de Microsoft, plusieurs applications et démonstrations ont tenté de combiner l'information de plusieurs caméras 3D afin de préciser les résultats obtenus. Les bibliothèques *Kinect SDK* et *OpenNI* ne supportent pas complètement de multiples caméras 3D, ce qui a grandement limité les efforts de fusion de donnée de squelettes.

En fait, il est possible d'obtenir les images de profondeur de plusieurs dispositifs dans une seule application, mais l'extraction d'un squelette distinct de chacune des images est impossible. Une solution de contournement à ce problème est d'instancier un processus séparé par dispositif. De cette façon, les squelettes peuvent être obtenus de façon indépendante et leur information est relayé à un serveur principal.

Ainsi, deux approches distinctes ont fait surface lors de l'acquisition de données de caméras 3D indépendantes. La première combine les données brutes des images de profondeur, générant un *nuage de points* qui est ensuite traité pour extraire l'information désirée. La seconde combine l'information squelettique obtenue indépendamment de chaque dispositif en tenant compte du point de vue de chaque caméra.

Finalement, l'utilisation de plusieurs caméras 3D amène un problème fondamental. Puisque les caméras 3D sont non seulement des capteurs, mais également des émetteurs, il risque d'y avoir interférence entre les différents signaux générés par les caméras 3D. En effet, chaque caméra émet un laser d'une certaine longueur d'onde en infrarouge proche. Cette longueur d'onde est la même pour chaque caméra. Lorsque les lasers infrarouges rebondissent sur une surface, ceux-ci peuvent s'entrecroiser et causer des interférences ; un capteur pourrait percevoir un rayon infrarouge qui n'est pas émis pas lui-même ou la collision des rayons résultent en signaux intraitables. Plusieurs cas d'utilisation de 2 caméras ou plus ont été reportés, mais mal documentés. Une étude Berger et al. (2011) a montré que l'utilisation de jusqu'à 4 caméras 3D *Kinect* ne cause pas suffisamment d'interférence pour être significatif sur le suivi de squelettes, surtout dût à la redondance d'information reliée à l'utilisation de multiples senseurs.

Fusion de données et recalage

Un nuage de points est une série de points dans un espace 3D. Dans le cas spécifique des caméras 3D, chaque pixel valide généré est un point 3D, dont les coordonnées sont relatives à la position du senseur. Ainsi, puisque chaque caméra possède sa propre référence, un point donné physique n'aura pas les mêmes coordonnées pour chaque senseur.

Pour pouvoir utiliser l'entrée de chaque dispositif, il est important de faire correspondre les données de chacun de ceux-ci afin de combiner les données. Cette procédure est appelée *recalage* ou *«registration»*. Le recalage peut être fait en 2D sur des images, ou en 3D sur un nuage de points. Aussi, le recalage peut être fait sur des points-clés de l'image 3D ou sur le squelette généré par le *Middleware* utilisé. L'avantage de l'utilisation du squelette généré est que les points-clés sont les jointures du squelette. Ainsi, l'identification des points-clés et la correspondance entre les images ne sont plus nécessaires.

Comme mentionné précédemment, la fusion de données peut-être faite sur les images de profondeur obtenues par les caméras 3D et interprétées comme nuages de points. Cette approche pose deux problèmes. Premièrement, elle nécessite l'identification de points-clés dans les images de profondeur du senseur primaire et leur correspondance dans les images générés par les senseurs secondaires. La position des senseurs n'étant pas connue, il pourrait être impossible de faire correspondre des points-clés de différents senseurs, par exemple, si les caméras sont une face à l'autre. Deuxièmement, l'importance de l'interférence des rayons infrarouges causée par l'utilisation de multiples caméras 3D demeure incertaine. Un rapport technique montre qu'il est possible d'annuler l'interférence entre deux caméras 3D Schröder et al. (2011) en alternant la projection du laser entre les deux dispositifs. Cette technique demande du matériel supplémentaire et est peu pratique. Une autre étude montre que l'interférence est négligeable avec moins de 5 caméras 3D sur le résultat des squelettes obtenus *Kinect* Berger et al. (2011), mais qu'il y a une grande interférence sur les images de profondeur individuelles. Ceci rend la fusion de données avant le traitement des images par le *Middleware* douteuse.

La seconde approche à la fusion de données est de l'appliquer suite au traitement intermédiaire des images de profondeur, c'est à dire après l'obtention des squelettes individuels de chaque dispositif. Une fois le recalage effectué sur chaque squelette, on obtient une série de positions dans le même référentiel et il devient alors facile de combiner les données obtenues.

L'idée du recalage est de faire correspondre des points dans chaque référence et de trouver une transformation qui minimise la distance (l'erreur relative) entre ces deux points. Le processus est répété jusqu'à ce que l'erreur tombe en dessous d'un seuil donné. Les transformations rigides sont des transformations dans un espace euclidien qui conserve les distances entre chaque point. Il est à noter que l'estimation de transformations rigides se fait par paires, c'est-à-dire que chaque senseur a sa propre transformation, par rapport au senseur principal. Des méthodes de recalage de plusieurs ensembles de données simultanées existent Huber and Hebert (2003), mais ne montrent pas d'avantages au recalage individuel. De plus, l'incertitude reliée à la collecte de données avec plusieurs caméras 3D peut introduire des erreurs dans le recalage automatique simultané de tous les senseurs.

L'estimation de transformations rigides indépendantes peut être implémentée de plusieurs façons. Une étude a été faite en 1997 comparant quatre algorithmes communs utilisés pour leur calcul Lorusso et al. (1997). Il est démontré que les techniques ne montrent pas de résultats significatifs quant à la précision de la transformation obtenue, mais que l'algorithme par SVD (*Singular Value Decomposition* ou *Décomposition par Valeur Singulière*) et UQ (*Unit Quaternions* ou *Quaternions unitaires*) sont plus stables, donc converge plus rapidement vers la solution. La solution par SVD montre la meilleure stabilité et précision globale. De plus, la technique par SVD montre un avantage de plus par rapport à l'estimation basée sur Levenberg Marquardt, qui est implémentée dans la bibliothèque PCL (*Point Cloud Library* PCL (2012)) puisqu'elle n'est pas itérative et ne demande pas d'estimation initiale. La SVD trouve la solution en une seule étape en faisant quelques suppositions initiales face au problème, qui n'ont pas de conséquences dans le cas présent.

2.3 Capture de mouvements

La capture de mouvement (*Motion Capture* ou *MoCap*) est un sujet très large qui intéresse plusieurs secteurs. Le cinéma, les jeux vidéo, la médecine, l'entraînement, la musique et de nombreux autres domaines sont intéressés par des solutions de capture de mouvements. Aussi, la capture de mouvements peut être appliquée autant au corps humain en entier qu'à une partie spécifique, comme le visage ou les mains ou même d'autres sujets comme des animaux.

La branche de capture de mouvement étudié est la capture de mouvement d'un corps humain complet. Il s'agit, dans la majorité des cas, de capturer le plus précisément possible les mouvements d'un sujet humain afin de les numériser et les utiliser dans un contexte particulier. Deux grandes familles de capture de mouvement existent pour le corps humain, soit des techniques envahissantes et des techniques sans marqueurs.

2.3.1 Techniques invasives

Les techniques de capture de mouvements envahissantes sont habituellement plus précises. Cependant, elles nécessitent que le sujet porte des senseurs, marqueurs et/ou accéléromètre. Toutes les données sont recueillies en temps réel et une animation est générée. Le but ultime habituellement visé par les méthodes envahissantes est de recréer le plus fidèlement les mouvements d'un sujet. Par contre, elles sont généralement très dispendieuses et le port d'équipements spécialisés peut encombrer les mouvements du sujet. Finalement, les animations générées, malgré qu'elles soient très précisent, nécessitent habituellement des retouches d'artistes lorsqu'elles sont utilisées dans des applications de multimédia ou dans le cinéma, car le sujet ne produit pas nécessairement le mouvement espéré.

Il existe deux grands types de captures de mouvements avec marqueurs. Le premier type utilise des points de références colorés afin de situer les points-clés dans l'espace. Cette technique utilise des caméras RGB à haute résolution afin de suivre les points-clés. Dans cette technique, le sujet sera également habillé de façon à augmenter le contraste entre lui et les points suivis, afin de mettre ceux-ci en évidence. Par exemple, trois points blancs peuvent être fixés à chaque joint d'un sujet habillé en noir. Plusieurs caméras observeront le sujet et permettront le suivi de la position et la rotation de chaque joint.

Le deuxième type de capture de mouvement avec marqueur utilise des accéléromètres, gyroscopes et un système de positionnement pour chaque joint du sujet. Les senseurs envoient directement leur information de positions et de rotations à un poste de travail. Cette technique a pour avantage de ne pas nécessiter que les joints du sujet soient dans le champ de vision d'une caméra.

2.3.2 Capture sans marqueurs

Les techniques de capture de mouvement sans marqueur (habituellement appelés *Markerless Motion Capture*) sont souvent basés sur des caméras. Aussi, depuis la venue de la caméra 3D *Kinect de Microsoft*, plusieurs solutions sont basées sur les caméras de profondeur.

Une technique largement utilisée avec des caméras RGB normales demande une calibration initiale, permettant de détecter l'arrière-plan. Ensuite, cet arrière-plan peut être soustrait des images nouvellement capturées afin d'isoler le sujet capturé. Cette technique a été jumelée à la capture sonore de chaque caméra utilisée pour la synchronisation des images, le suivi de trait et la densité de détails de l'arrière-plan afin de stabiliser les résultats obtenus Hasler et al. (2009). Le squelette du sujet capturé est ensuite extrait des images capturées par la caméra RGB et on peut en extraire les positions de chaque joint dans l'espace. Par contre, puisque les caméras RGB ne fournissent aucune information de profondeur sur l'image capturée, cette technique souffre des mêmes limites de précisions quant à la position des joints dans l'axe de profondeur.

La technique précurseur au suivi avec des caméras 3D utilise un modèle articulé du corps humain afin de préciser la capture de mouvement. Cette technique offre une façon facile d'extraire le squelette des images de profondeur fournies par la caméra 3D. Une approche utilisant cette technique consiste à générer une représentation voxel du modèle articulé à partir des images Sundaresan and Chellappa (2005), qui a l'avantage d'être adaptable selon le niveau de précision requis grâce à la nature évolutionelle des voxels. Par contre, la limite de la précision de la capture est toujours déterminée par la résolution en profondeur de la caméra utilisée.

Captures à bases de caméras de profondeur

Depuis l'apparition de la caméra *Kinect de Microsoft*, les caméras 3D sont devenues un outil essentiel à la capture de mouvements sans marqueurs. Non seulement les caméras 3D sont maintenant peu coûteuses (environ 150$), mais les *Middleware* et kits de développement compris avec ces dernières fournissent un squelette complet du sujet Shotton et al. (2011). En utilisant une combinaison des suivis de squelettes intégrés aux caméras 3D et d'autres techniques de fusion, plusieurs recherches et exemples de capture de mouvements ont été faits.

Le *Flexible Action and Articulated Skeleton Toolkit* (*FAAST*) permet l'intégration facile de squelettes conformes à *OpenNI* à des jeux et des applications de réalité virtuelle. Un serveur VRPN (*Virtual-Reality Peripheral Network* fournit une interface simple pour accéder au squelette généré et mapper en temps réel des mouvements à des touches du clavier. De plus, FAAST fournit un suivi de l'orientation de la tête du sujet à l'aide de techniques de vision par ordinateur Suma et al. (2011).

2.4 Cinématique Inverse

Les équations de la cinématique inverse ont été premièrement développées afin de permettre aux robots-ouvriers de calculer l'orientation de leurs membres afin d'atteindre leur position finale. Les applications de systèmes de cinématique inverse en infographie sont souvent limitées aux jeux vidéo, où les animations ne peuvent pas être entièrement prédéterminées. En utilisant la position d'une cible comme référence, la cinématique inverse permet de calculer la position de chaque joint du squelette dans l'ordre inverse : en commençant des extrémités vers le *Root Node*, soit parent de tous les autres joints.

Les limites d'angle entre les os et les degrés d'influences de chaque cible sur les os du système sont une autre composante clé des systèmes de cinématique inverse modernes. Les limites d'angles empêchent le squelette étudié de prendre des poses qui seraient impossibles par un corps similaire en réalité. Par exemple, si l'on déplace la cible de la tête d'un sujet humain derrière celle-ci, les limites du modèle de cinématique inverse empêcheraient la tête de tourner à 180 degrés.

Les degrés d'influences permettent de définir l'importance d'une cible par rapport au déplacement qu'elle peut engendrer sur les autres joints du corps étudiés. Par exemple, si le poignet d'une sujet humain est attribué un très grand degré d'influence, le déplacement de la cible du poignet pourrait influencer le positionnement de tout le tronc et même les pieds du sujet.

2.4.1 Human IK

La bibliothèque de cinématique inverse appliquée au corps humain d'*Autodesk*, appelée *HumanIK* ou *Human IK* Autodesk (2012), implémente les calculs nécessaires pour faire de la cinématique inverse. Elle définit également un modèle squelettique du corps humain et lui attribue des limites pour chaque joint. De plus, elle attribue les influences que chaque joint peut avoir sur les autres dans le modèle du corps humain. Entre autres, le modèle du corps humain implémenté dans la bibliothèque a été perfectionné afin de simuler avec plus de précision des mouvements sportifs dans les jeux vidéo. Dans ce scénario, une animation de base est appliquée à un modèle 3D et corrigée à l'aide de la bibliothèque HumanIK.

Par exemple, dans un jeu de hockey sur glace, le gardien de but possède une animation permettant de prévenir un lancé visé dans le coin gauche du filet. Par contre, il sera extrêmement rare, voire impossible, qu'un joueur lance la rondelle exactement à l'endroit qui coïncide avec le mouvement du gardien de but. Dans un tel scénario, la bibliothèque HumanIK est donc utilisée. L'effecteur associé à la main du gardien qui doit faire l'arrêt est placé à la position de la rondelle, et un poids important lui est associé. Ce poids détermine l'influence que peut avoir l'effecteur sur le reste du squelette et du modèle de l'être humain. Ainsi, le résultat est que l'animation de base sera influencée par l'effecteur et le gardien de bût, sa position le permettant, effectuera un arrêt.

Modèle de Joints

Le modèle humain dans la bibliothèque *HumanIK* d'Autodesk permet de définir au-delà de 150 joints. Chacun de ceux-ci ont des degrés de liberté variables dépendants de leur positionnement et de leurs relations avec leurs parents. Ainsi, les joints du tronc du sujet ont moins de liberté que ceux du coude par exemple, car ils sont directement liés aux autres joints du tronc, donc ne peuvent pas bouger indépendamment de ces derniers.

Chaque degré de liberté de chaque joint est directement relié à sa liberté dans le corps humain. C'est en examinant précisément, à l'aide de plusieurs spécialistes, les limites du corps que Autodesk a su incorporer les angles limites d'un humain. De plus, la bibliothèque permet d'incorporer des degrés de flexibilité de chaque joint afin de simuler plusieurs humains avec des degrés de flexibilités variants, tels que des gymnastes ou des personnes âgées.

Malgré sa robustesse et fiabilité par rapport aux poses individuelles du corps humain, la bibliothèque Human Inverted Kinematics d'Autodesk ne définit aucune contrainte spatiale ou temporelle quant aux mouvements subséquents. Ainsi, deux poses calculées de façon séquentielle peuvent être complètement différentes.

2.5 Recherches concurrentes

Naturellement, la popularité des caméras 3D en fait l'objet de plusieurs recherches. Une recherche est particulièrement orientée vers les utilités pratiques et la vision des caméras Kinect en dehors des applications dans le domaine du multimédia Zhang (2012).

Une approche similaire à la recherche proposée tente de réconcilier la capture de mouvement par caméra 3D avec un modèle physique d'un être humain plutôt qu'un modèle basé sur la cinématique inverse Shum and Ho (2012). Cette approche montre beaucoup d'intérêt, car elle a le potentiel d'être complémentaire à la technique proposée et les deux peuvent être combinés afin d'améliorer globalement les résultats. Les résultats de la recherche basée sur un modèle physique montrent particulièrement une grande amélioration des pics de déplacements dans les séquences capturés. Ceci résulte en des mouvements beaucoup plus naturels et une animation plus fluide, malgré que la précision de la capture ne montre pas d'amélioration significative.

Cette technique montre un intérêt particulier par rapport à la technique proposée dans ce mémoire. Par sa nature physique, la recherche Shum and Ho (2012) a le potentiel d'être intégrée avec la cinématique inverse afin d'améliorer davantage les résultats.

CHAPITRE 3 MÉTHODOLOGIE ET RÉSULTATS COMPLÉMENTAIRES

3.1 Introduction

La méthodologie établie sert de guide au travers de l'expérimentation afin de valider l'hypothèse émise et de vérifier les conclusions tirées de la recherche. Celle-ci vise à définir quels éléments doivent être développés afin de répondre aux besoins de l'expérience. Elle permet également de mesurer la progression de l'étude et de se concentrer sur les éléments importants soulevés.

3.2 Méthode scientifique

3.2.1 Hypothèse

L'hypothèse présentée par cette recherche est que l'animation résultante de la capture de mouvements à l'aide d'une caméra de profondeur peut être significativement améliorée en combinant l'analyse squelettique de plusieurs caméras de profondeur avec des techniques de cinématique inverse appliquée au corps humain. Plusieurs éléments sont impliqués dans cette hypothèse.

Tout d'abord, cette hypothèse implique que la combinaison des données provenant de plusieurs caméras 3D est possible et que leur qualité est suffisante pour utiliser au moins deux de ces dispositifs simultanément. Aussi, l'hypothèse implique que la combinaison des squelettes résultants de chaque caméra individuelle améliorera la moyenne de précision de chaque pose individuellement capturée. Il est également attendu que les transitions entre chaque pose individuelle soient améliorées. Ces facteurs contribueront enfin à une animation finale plus précise et de meilleures qualités subjectives.

La seconde implication de l'hypothèse est que l'utilisation d'un système de cinématique inverse, tel que la bibliothèque HumanIK d'Autodesk, mènera également à des séquences d'animation de meilleure qualité que celles créées simplement avec la caméra 3D. L'intégration et l'utilisation du système de cinématique inverse dans le système de capture devront corriger des défauts visuels présents lors de la capture par simple caméra 3D. Le résultat sera une animation de qualité supérieure, débarrassée d'artéfacts liés aux résultats de la capture par caméra 3D, tel que des positions impossibles pour un être humain.

3.2.2 Expérimentation

Afin de répondre aux questionnements que présente la recherche, une démarche rigoureuse est mise en place afin de recueillir les données nécessaires pour tirer les bonnes conclusions par rapport aux hypothèses émises. Chaque implication de l'hypothèse sera sujet d'une expérience indépendante et ensuite intégré à la solution globale proposée. Une application de capture de mouvement sera ainsi construite. L'implémentation de la capture de mouvement avec une seule caméra 3D permettra d'avoir une référence afin de comparer les résultats de chaque élément ajouté. Des fonctionnalités permettant d'enregistrer et de rejouer des séquences de captures particulières sont aussi nécessaires afin de permettre de répéter les mêmes séquences sous différentes conditions. Chaque aspect de la recherche y sera ensuite ajouté individuellement, et intégré à la solution globale, avec un mécanisme permettant de facilement activer ou désactiver chaque fonction.

La première étape de l'expérimentation consiste à déterminer la limite du nombre de caméras 3D que l'ont peut utiliser simultanément. Dût aux interférences entre elles, ajouter des Kinects pointées au même sujet diminue la qualité des images produites. Certaines études ont déjà été faites sur les limites de l'utilisation de multiples caméras Kinects Berger et al. (2011), mais puisque les conditions d'éclairage, les sources de lumières ambiantes et la texture des vêtements du sujet peuvent avoir une grande influence sur les résultats, ces expériences seront répétées dans les conditions de capture de mouvement particulières à notre expérience. La limite de caméra pouvant être utilisée simultanément sera déterminée par la capacité à détecter le squelette du sujet avec le bruit produit par les autres caméras. Par exemple, si un squelette est détecté avec 4 caméras, mais pas avec 5, peu importe la qualité de la détection, la limite sera fixée à 4.

Afin de vérifier l'effet de l'utilisation de multiples caméras 3D simultanément, un module permettant de lancer et enregistrer des séquences de plusieurs caméras 3D simultanément devra être implémenté. Dût aux limites établies par les bibliothèques de OpenNI et KinectSDK, une architecture client-serveur sera implémentée afin de contrôler multiples caméras 3D simultanément. Chaque client, lancé sur une machine indépendante, aura le contrôle d'une caméra et sera en mesure de lancer l'algorithme d'extraction de squelettes à partir de l'image de profondeur. L'information de chacune des caméras est ensuite recueillie par le serveur, qui sera en mesure de traiter les données de façon adéquate.

Une fois recueillie, l'information squelettique provenant de chaque caméra doit être combinée. À cette fin, la bibliothèque PCL sera utilisée. Leur algorithme de transformée rigide permet d'estimer la matrice de transformation d'un espace de référence à un autre. Ainsi, en utilisant une caméra 3D comme référence, il sera possible de transposer les points recueillis de chaque caméra dans l'espace de la caméra de référence. Suite à cette transformation, il sera alors possible de directement comparer les articulations de chaque squelette. Une fois la transformation estimée et l'information de chaque squelette recueilli par le serveur, celui-ci saura filtrer les données erronées et obtenir un squelette moyen en effectuant une moyenne de chaque squelette individuel.

Avec les données recueillies, l'étape suivante sera de déterminer l'effet des multiples caméras sur la qualité des poses recueillies. Ayant précédemment établit une limite supérieure au nombre de caméras 3D pouvant être utilisé simultanément, il faudra déterminer une limite pratique. Celle-ci sera déterminée en capturant des séquences similaires avec une, deux, trois, etc. caméras, jusqu'à la limite déterminée. Ensuite, ces séquences seront analysées afin de quantifier la qualité des squelettes combinés en utilisant un nombre variable de caméras 3D.

La prochaine étape sera d'intégrer les concepts de cinématique inverse et les fonctionnalités de la bibliothèque Human Inverted Kinematics d'Autodesk à la solution de capture de mouvement proposée. Celle-ci demande une calibration initiale du personnage poussé par sa technologie. C'est pourquoi le sujet devra débuter chaque séquence avec une pose de calibration en «T-Stance», soit les bras étendus horizontalement à la hauteur des épaules, paumes face au sol, avec les pieds faisant face à la caméra de référence, à la largeur des épaules. Cette pose permet d'initier les proportions et les limites qu'utilise la bibliothèque HumanIK. Ensuite, afin de résoudre la position finale du personnage, la bibliothèque demande une pose initiale en entrée, ainsi que des effecteurs pour chaque joint. Chaque effecteur a une influence sur l'articulation qui lui est attribuée, mais peut également influencer le reste du squelette selon les contraintes qui lui sont appliquées. Par exemple, l'effecteur du poignet gauche, si tiré jusqu'au sol, peut causer les genoux du personnage à fléchir afin d'atteindre son bût.

Il y a deux façons de procéder à l'intégration de la bibliothèque HumanIK à la solution de capture de mouvement. La première consiste à utiliser la pose capturée dans le dernier cadre comme entré pour HumanIK. Le dernier squelette sera donc la référence pour résoudre la pose finale. Avec cette approche, l'information squelettique recueillie par le serveur sera alors utilisée pour placer les effecteurs du système de cinématique inverse, qui auront une grande influence sur le squelette. La pose ainsi résolue sera alors utilisée comme clé d'animation, et servira d'entrée pour la prochaine itération.

La seconde méthode d'intégration de HumanIK dans le système de capture de mouvement utilise la cinématique inverse pour corriger les poses extraites directement des caméras 3D. En utilisant directement le squelette capturé comme entrée et pour placer les effecteurs, l'étape de résolution de HumanIK est utilisée afin de vérifier la pose obtenue. Cette approche diffère de celle élaborée plus haut, car elle n'utilise aucunement les données calculées précédemment.

Les deux méthodes d'intégration seront implémentées dans l'application de capture de mouvement. Elles seront appliquées indépendamment sur les mêmes séquences de captures afin d'être analysées objectivement.

Finalement, d'autres composantes seront ajoutées afin de parfaire l'animation capturée. Les fonctions à ajouter seront déterminées par les limites rencontrées avec l'intégration de la cinématique inverse au système de capture de mouvement. Par exemple, l'incapacité de capturer les mouvements des mains ou l'orientation de la tête du sujet sera sans doute des potentiels des fonctions à ajouter.

Afin d'augmenter la qualité des animations finales, il sera aussi nécessaire de simuler les captures de certaines parties du corps, telles que les mains et la tête. Puisque les caméras 3D ne supportent pas la détection de l'orientation de la tête et de l'état des mains du sujet, les animations résultantes semblent rigides et irréalistes. Il sera donc important d'interpréter les images 3D afin de trouver cette information ou de les simuler. Ainsi, l'état des mains sera recherché dans les images 3D produites par la caméra centrale. À cause de la résolution limitée et l'erreur associée à l'information de profondeur, la position de chaque doigt ne peut pas être extraite. Par contre, l'état de chaque main (ouverte ou fermée) pourrait être extrait. Un module qui tentera de suivre la position de chaque main et d'en extraire l'état sera donc implémenté dans l'application finale. Ensuite, des contraintes de déplacement temporelles et la cinématique inverse seront appliquées au doigt afin de simuler leurs mouvements.

Finalement, des filtres de mouvements seront appliqués à l'animation produite afin de réduire les sauts de positions résultants de l'imprécision des caméras 3D. Les bibliothèques de capture de squelette intègrent un filtre basé sur la technique de Holt Holt (2004); Winters (1960), mais celle-ci réduit considérablement la résolution et la précision des mouvements capturés. Par contre, les grands sauts de positions des membres du sujet sont la plus grande cause de l'imprécision de la capture finale. Ainsi, un filtre sera implémenté afin de limiter les mouvements de chaque membre d'une trame à l'autre. Cette approche filtrera les très grands déplacements rapides des membres, tels que des sauts impossibles de position, tout en laissant intacte la précision des mouvements.

Chaque composante de l'application et chaque étape de l'expérimentation seront évaluées indépendamment afin de bien comprendre l'effet de chaque composante sur le résultat final. De plus, les composantes retenues indépendamment seront mesurées ensemble afin de valider qu'elles n'interfèrent pas entre elles. Aussi, il est même possible que les composantes synergisent entre elles et donnent un résultat encore meilleur que prévu.

3.2.3 Analyse de données

Les résultats recueillis sont analysés de façon indépendante pour chaque composante du système créé dans le cadre de la recherche. Le but de la recherche est d'améliorer les animations résultantes de la capture de mouvement à l'aide de caméras 3D. Ainsi, chaque aspect de la recherche doit améliorer quantitativement ou qualitativement les résultats.

Le premier module à implémenter afin de permettre une analyse rigoureuse des données est un système d'enregistrement des animations capturées. Ensuite, un module de lecture est implémenté afin de permettre de rejouer les enregistrements. Avec ce système en place, il est alors possible de rejouer les mêmes séquences de mouvements en activant et ajustant différentes composantes du système.

Les caméras 3D sont une nouvelle technologie et leur précision laisse à désirer. Ainsi, pour que leur utilisation soit justifiable dans une application pratique, la précision des captures doit être améliorée. C'est pour cette raison que l'évaluation de la précision de la méthode de capture de mouvements suggérée est faite en comparant les mouvements obtenus avec cette technique à ceux obtenus avec un système commercial dont la précision est suffisante pour des applications pratiques et bien documentée.

La comparaison est donc faite avec un système *Flock Of Birds* (FOB) de Ascension. Celui-ci a une résolution de 0.8mm avec une erreur de 2.5mm, ainsi qu'une fréquence de rafraîchissement de 100Hz. Le prix d'un tel système est de l'ordre de 6000$, environ 13 fois le prix de 3 caméras Kinect ! Afin de comparer les mêmes données, les séquences de mouvements d'un sujet ont été capturées simultanément avec le système FOB et avec la technique de capture de mouvement suggérée dans cette recherche. De plus, afin d'éliminer le plus de bruit et de variables qui peuvent influencer les résultats, un mouvement simple a été utilisé. Seul le bras droit du sujet est en mouvement. Ce mouvement simple a le double avantage de garder le mouvement simple, tout en gardant une implication de 3 joints distincts, soit l'épaule, le coude et le poignet du sujet.

Ensuite, chaque module implémenté est analysé de façon indépendante afin de déterminer leur effet sur la précision et sur la qualité de la séquence de mouvements produite. En utilisant la capture de mouvement d'une seule caméra 3D avec aucun traitement comme référence, il est alors possible de mesurer la différence de précision, en comparant chaque capture avec les mesures produites par le système FOB.

La précision des mouvements est évaluée en comparant les graphiques des mouvements dans le temps de chaque joint, sur chaque axe de translation. Ainsi, en mesurant les mouvements de l'épaule, du coude et du poignet du sujet à l'aide du système FOB, nous obtenons trois

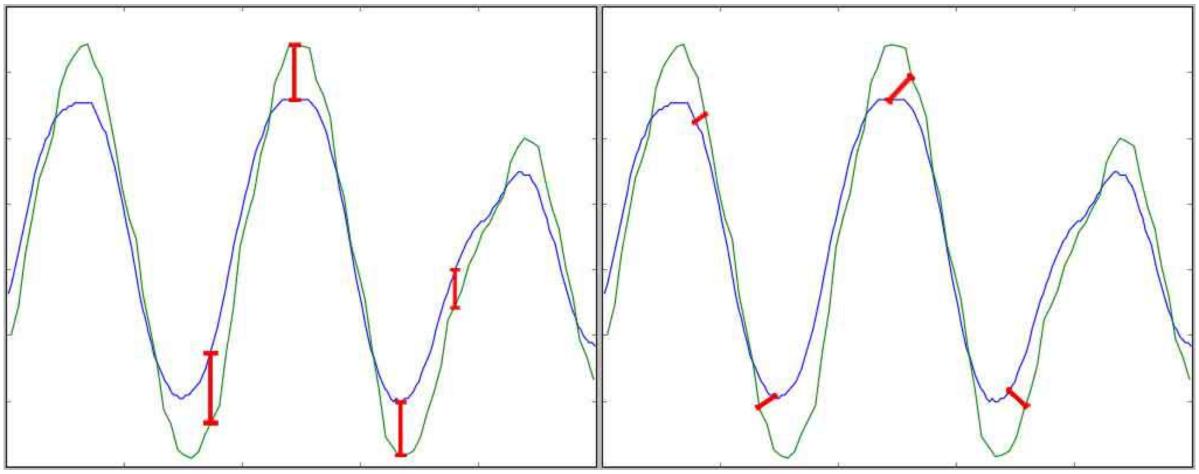

Figure 3.1 Une courbe des mouvements de poignets dans l'axe des Y (hauteur) montrant les deux mesures pertinentes pour comparer les graphiques. À gauche : La différence absolue entre chaque point correspondant. À droite : La plus petite distance possible entre les courbes.

points de données pour chaque trame capturée. Les graphiques produits par le système FOB et la méthode de capture proposée sont superposés en utilisant l'horodatage de chaque point de donnée (Figure 3.1).

Les courbes sont comparées à l'aide de quatre indicateurs clés (Figure 3.1). Le premier indicateur consiste à calculer la distance absolue entre chaque point de donnée correspondant dans les courbes comparées. Cette mesure permet de calculer la limite supérieure de la précision associée à la technique de capture de mouvement suggérée. Elle permet également de calculer directement la précision absolue de la courbe d'animation produite par la capture de mouvement par caméra 3D. Par contre, cette mesure est très sensible aux pics et aux déviations par rapport à la courbe de référence. Puisque qu'il est bien connu que la capture de mouvement par caméra 3D produit beaucoup de variances par rapport au mouvement réel, la valeur de la mesure par différence absolue augmente rapidement en fonction des déviations. Cette mesure peut être utilisée directement pour comparer la précision des mouvements, car une plus petite valeur indiquera forcément une meilleure précision dans la capture.

La deuxième mesure fait abstraction de l'horodatage des données dans le graphique. Elle consiste a mesurer la plus petite distance entre un point donné de la courbe produite par la capture par caméra 3D et n'importe quel point sur la courbe de référence. Cette méthode du point le plus proche permet de détermine la limite inférieure de la précision de la capture de mouvement. De plus, elle n'est pas aussi sensible aux pics et aux déviations par rapport à la courbe de référence en plus d'ignorer les distorsions dans le temps dû à la vitesse de capture des caméras 3D. Une plus petite valeur ne signifie par nécessairement une meilleure précision de capture, mais est un indicateur de la qualité visuelle de l'animation par rapport au vrai mouvement.

Deux autres mesures sont également prises. Premièrement, la différence de pics dans la courbe est enregistrée. Un pic est défini comme étant une déviation de plus de 10 centimètres sur 0.13 seconde (4 trames à 30 trames par secondes). En comparant la différence de pics dans la courbe de référence au nombre obtenu dans la courbe produite par la capture par caméra 3D, nous obtenons le nombre d'imperfections produites par notre technique de capture. Limiter le nombre de pics produira forcément des captures de meilleure qualité.

Deuxièmement, le nombre de déviations entre la courbe de référence et la courbe de la capture est compté. Une déviation se produit lorsque la forme de la courbe de référence ne correspond pas à la courbe observée. Par exemple, si la courbe du système FOB est concave entre t=1.0 et t=2.0 et que la courbe mesurée entre ces temps est convexe, il s'agit d'une déviation.

Ces quatre mesures permettront ainsi de déterminer les bornes inférieures et supérieures de la précision de notre technique de capture, ainsi que de mesurer la précision et la qualité de l'animation produite.

3.2.4 Vérification de données

Les données recueillies durant les captures de mouvements et analysées par la suite doivent être vérifiées afin de valider leur fidélité. Pour ce faire, les captures de mouvements simultanées avec le système FOB sont répétées et analysées de façon indépendante. Ensuite, la moyenne des valeurs obtenues sera extraite afin d'obtenir des résultats fiables.

Pour pouvoir comparer la grande quantité de graphiques produits par chaque capture et recueillir les quatre mesures recherchées, des scripts sont écrits en Python afin d'automatiser le processus. En automatisant le processus d'analyse de résultats, nous pouvons obtenir et analyser une très grande quantité de courbes d'animations en très peu de temps.

Ainsi, quatre mouvements seront enregistrés et répétés trois fois chaque. Le premier mouvement demande au sujet de bouger son bras étendu de l'avant vers l'arrière, en limitant les mouvements en hauteur. Le second consiste à effectuer un mouvement similaire, mais de haut en bas, afin de limiter le déplacement dans l'axe faisant face au sujet. Le troisième vise à limiter les mouvements latéraux et demande au sujet de faire des cercles avec son bras étendu. Finalement, le quatrième mouvement demande au sujet d'effectuer un mouvement de coup de poing, afin de vérifier les effets de la vélocité dans la capture.

De plus, chaque mouvement sera reproduit avec un nombre variable de caméras 3D, commençant avec une seule et augmentant jusqu'à ce qu'on atteigne la limite de caméras déterminée durant l'expérimentation initiale. Si cette limite est, par exemple, de quatre caméras, nous obtiendrons un total de 192 graphiques, soit quatre mouvements représentés par trois courbes de translation dans le temps (une pour chaque axe), répétés quatre fois, avec quatre variations du nombre de caméras 3D.

Après avoir déterminé la validité de chaque composante du système proposé en mesurant et vérifiant nos mesures de précision et de qualité des animations produite, nous vérifions que les composantes ne produisent pas d'interférences entre elles. De la sorte, chaque composante montrant une amélioration dans nos mesures sera jumelée entre elles et mesurée à nouveau. Finalement, toutes les composantes qui montrent une amélioration sont également activées simultanément afin de mesurer l'effet du système au complet sur la précision et la qualité de l'animation finale produire par la technique de capture suggérée.

Afin de vérifier qualitativement les résultats obtenus par la méthode de capture de mouvement proposée, deux tests subjectifs sont passés pour chaque composante du système. Le but étant de produire une capture non seulement plus précise, mais également visuellement plus attirante, il est important d'obtenir des opinions diverses. Pour chaque composante, les captures obtenues sont appliquées à des modèles 3D et compilées dans de courts vidéos. Ces captures sont ensuite présentées à un groupe d'individus qui indiquent, à l'aveuglette, c'est-à-dire sans savoir quel vidéo est associé à quelle technique, quelle capture ils préfèrent ; celle avec ou sans la composante additionnelle activée. Même si une composante améliore la précision d'une capture, elle peut y amener des artéfacts visuels qui détériorent la qualité globale de la capture. Le contraire est également possible. Même si une composante diminue la précision de la capture, si celle-ci améliore la qualité visuelle du résultat final, elle peut être retenue dans la solution finale. Le groupe d'individus pour participer au test à l'aveuglette est un groupe de 66 étudiants suivant un cours d'infographie.

Finalement, un expert dans le domaine sera également consulté afin de valider la performance et la qualité de la solution proposée, ainsi que de valider son intégration dans un pipeline de production dans une application pratique. Plusieurs animateurs professionnels sont engagés par Autodesk et seront disponibles à des fins de consultations.

3.3 Intégration commerciale

Nous avons déjà établi le potentiel de la recherche à être intégré à une solution commerciale de capture de mouvement. La méthodologie décrite ci-haut renforce l'utilisation pratique de la recherche en y amenant un cadre bien défini qui permettra de bien identifier quelles composantes ont le potentiel d'être utiles dans un contexte de production.

Étant donné la collaboration avec la compagnie Autodesk inc., les méthodes de capture de mouvements suggérées ont le potentiel d'être directement intégrées dans un ou plusieurs logiciels de création de contenus multimédias de la compagnie. Entre autres, le logiciel *Motion Builder* se spécialise dans la capture de mouvement. Une solution de capture de mouvement sans marqueurs fiable serait un atout majeur dans une prochaine itération du logiciel.

Avec des résultats significatifs, la recherche proposée pourrait être la fondation pour une solution de capture de mouvement fiable et abordable pour des studios de jeux. Ainsi, la capture de mouvement deviendrait plus accessible aux studios indépendants sur un budget limité. Chaque composante identifiée précédemment aura le potentiel d'être intégrée de façon indépendante, selon le contexte pratique de l'application.

Puisque la technique de capture de mouvement suggérée n'exige aucun marqueur sur le sujet, celle-ci est moins intrusive pour ce dernier. Ceci a comme conséquence de produire des mouvements plus naturels et fluides puisque l'acteur ne sera pas encombré par des marqueurs qui peuvent gêner ses mouvements. La complexité des installations nécessaires à la capture de mouvements sera aussi grandement diminuée, ce qui réduira considérablement le temps nécessaire pour effectuer une capture complète. Ainsi, une capture peut être refaite plusieurs fois à différents moments sans engendrer de coût supplémentaire, et permettra une itération rapide des captures nécessaires à la création de contenus multimédias.

Tous ces aspects font de la capture de mouvement par caméra 3D une solution idéale. En améliorant la qualité de ce type de captures, elles ont le potentiel de remplacer toutes les techniques présentement utilisées.

CHAPITRE 4 PRÉSENTATION DE L'ARTICLE

Ce chapitre présente l'article du prochain chapitre et résume la méthodologie utilisée pour atteindre les objectifs spécifiques décrits précédemment.

Nous décrierons premièrement les démarches et l'installation dédiées à combiner les données de plusieurs caméras 3D afin d'obtenir un squelette de base plus précis avant l'intégration de la cinématique inverse. Ensuite, les différents essais d'intégration seront détaillés et leurs résultats quantifiés et expliqués. Enfin, les méthodes de recherche d'information additionnelle afin de complémenter les séquences capturées seront expliquées.

Une fois les résultats compilés et présentés, nous les comparerons aux séquences de mouvements produites par simples caméras 3D, et mesurerons leur précision à l'aide d'un système de capture précis.

4.1 La capture par caméra de profondeur

La capture par simple caméra

La qualité d'une capture par caméra de profondeur est telle qu'elle n'est pas pratique pour des fins autres que de divertissement. La technique proposée recherche donc à augmenter sa précision de façon à amener la qualité des captures produites à un standard d'une qualité acceptable afin d'étendre leur utilité dans un contexte de production d'animation.

La capture par multiples caméras

On peut simplement penser que l'ajout de caméras de profondeur peut potentiellement mener à de meilleurs résultats de séquences de mouvements. Toutefois, par la nature active des capteurs de profondeur, l'interférence entre multiples caméras ne peut être négligée. Il est donc important de mesurer l'impact de l'utilisation de plusieurs de ces caméras simultanément.

L'approche

L'article de la section suivante décrit donc l'approche utilisée afin de poursuivre la méthodologie établie dans la section précédente. Chaque section de l'article décrit une composante de la technique de capture de mouvement proposée et soulève les problèmes rencontrés. Ensuite, l'article résume les résultats et en fait une analyse quantitative ainsi que qualitative afin de confirmer ou rejeter l'hypothèse émise.

4.2 Contenu de l'article

4.2.1 Capture du squelette

La première partie de l'article vise à décrire comment l'information squelettique provenant de multiples sources est combinée. Les limites matérielles et logicielles sont exposées et contournées afin de permettre de profiter de la redondance d'information venant de différentes sources. Le squelette résultant est enfin utilisé comme base pour le reste de l'intégration de la technique proposée.

4.2.2 Intégration de HumanIK

Le cœur de la recherche est axée vers l'intégration de la bibliothèque HumanIK dans le pipeline de production d'animation par caméra 3D. Dans cette section de l'article, les différents essais d'intégrations sont élaborés, mesurés et comparés. Les résultats sont ensuite présentés et une décision est prise sur l'effet bénéfique ou nocif sur la qualité de la séquence produite.

4.2.3 Informations supplémentaires

Les séquences produites après l'intégration de HumanIK montrent certaines déficiences qualitatives. Cette section de l'article vise à décrire comment de l'information supplémentaire a été dérivée des images de profondeur afin de profiter des capacités de la bibliothèque HumanIK. Ainsi, de l'information pour compléter le modèle humain, particulièrement l'état des mains et de la tête, est extraite et intégrée dans la solution de capture de mouvement par caméra 3D proposée.

CHAPITRE 5 ARTICLE 1 : MOTION CAPTURE TECHNIQUE INTEGRATING INVERSE KINEMATICS WITH MULTIPLE DEPTH CAMERAS

David Ménard[1], Benoît Ozell[1], and André Foisy[2]

[1] Département de génie informatique et génie logiciel, École Polytechnique de Montréal, Montréal QC, Canada

[2] Autodesk inc.

Abstract

In this paper, we propose a motion capture technique using multiple depth-cameras in combination with inverse kinematics to qualitatively improve on the animation produced by a single depth camera.

The method introduces independent components that are incorporated into the proposed technique, and evaluated separately. We retain the components which show an improvement, then combine and measure the final produced animations. We start by looking at the effects of using multiple cameras on the IK integration.

The results, validated by comparison with a precise motion capture system and a series of blind tests in which 66 volunteers participated, show a significant increase in animation quality. Interestingly, the measurements have shown an increase in both precision and quality after integrating IK to the depth camera data. This technique provides a precision improvement of about 30% as well as a significant visual improvement confirmed by all 66 of blind test volunteers.

Keywords: animation, human inverse kinematics, HumanIK, Kinect, depth-camera, HIK, motion capture, mocap, markerless

Submitted to: *Computer Graphics and Applications*, November 2014.

5.1 Introduction

The debut of Microsoft Kinect camera brought the depth-camera to an affordable price, triggering a multitude of demonstrations and applications using the depth sensor as an input device. Though the Kinect camera prevails as a gaming device, it has the potential to help game developers throughout production.

The Microsoft Kinect SDK and PrimeSense middlewares remain a relatively robust way to track users at an interactive rate, but their limitations become apparent when looking directly at the generated skeleton. In fact, from the middleware's point of view, there are no limitations to what the human body can do; Bones can bend at impossible angles, change length and move at super human speeds during the capture. This can allow for some error compensation and smoother movements in applications which do not require precision, but might not be the best approach for motion capture.

For these reasons, motion capture using depth-cameras has been limited. Although some efforts have been made to combine the data generated by multiple depth-cameras and filter noise to produce a clearer image from the Kinect (Bailey and Bodenheimer, 2012), these solutions are still susceptible to the skeleton tracking middleware limitations previously mentioned. In addition to this, multiple Kinect cameras interfere with each other, resulting in unstable images to work with.

Furthermore, motion capture data usually requires rotation information to be useful in production. Unfortunately, the Kinect SDKs rotational information is extremely limited and impractical in a motion capture setting. This is also true for practically all markerless motion capture techniques. Top of the line MoCap solutions will of course offer joint rotation data, captured via various combinations of accelerometers, gyroscopes and three-point references, but these solutions are often costly and cumbersome. They are not only expensive, but also require elaborate setups and artists will almost always want to retouch the generated animations to their liking.

In this paper, we propose an alternative to top grade commercial motion capture techniques by utilizing multiple Kinect cameras and an inverse kinematics solution applied to the human body to generate realistic animations. The goal is not to produce the most precise motion capture data possible, but to produce the best visually appealing and realistic animation that satisfactorily follows the subject's movements from a human body perspective.

We start by increasing the precision of the skeleton coordinates by combining the skeletons generated by simultaneously recording cameras. Using Point Cloud Library (PCL) (PCL, 2012), we can estimate a matrix to transform each skeleton into the target cameras coordinate space.

Following a T-Stance calibration, can integrate inverse kinematics using Autodesk's HumanIK solution (Autodesk, 2012). We do this by targeting the input skeleton onto an HumanIK skeleton definition that will drive the following animation frames. Each new input frame sets the positions for key HumanIK effectors which are used to solve the new pose, including rotations for each joint.

Additional steps are taken between each frame to increase the realism of the generated skeleton and take full advantage of the HumanIK library: skeleton bone sizes are limited within a certain threshold, joint movements normals are limited from frame to frame, head rotation is adjusted and hand states are detected, all mapped into the inverse kinematics skeleton. The resulting animations are smoother, more appealing and more human-like than the simple motion capture provided by a single depth camera.

The following paper describes the multiple aspects of the proposed motion capture technique. Section 5.2 takes a look at our approach to integrating multiple Kinect devices together to fuse the skeleton data in a way that can be used with the HumanIK library. Section 5.3 focuses on the different attempts at integrating the Kinect skeletons to a human inverse kinematics model. Section 5.4 elaborates on how we obtain additional data to further enhance the generated animation and how it is integrated to take advantage of the inverse kinematics system. Section 5.5 explains different filters, which are applied to smooth out the animations. Finally, Section 5.6 goes through the results of both the accuracy measurements and qualitative analysis.

5.2 Capturing the Skeleton

The skeleton produced by the depth-camera drivers and middleware is the foundation of the proposed motion capture technique. Although PrimeSense has yet to reveal its tracking algorithm, Microsoft has published their first tracking algorithm (Shotton et al., 2011), which uses Bayesian probabilistic decision making, a common machine learning approach in computer vision. They both present similar results and confidence levels (Hinchman, 2011). Although Kinect SDK supports predictive tracking, it can lead to many false positives that are not present in the PrimeSense implementation. In addition, PrimeSense has implemented OpenNI, a framework that allows for easy access to depth-camera capabilities and PrimeSense features. Through OpenNI, PrimeSense provides easy recording and playback systems, which are essential to reproducing and comparing results, and allows for easy post-processing of the rendered animation.

Due to both hardware and software limitations, only one depth-camera can be activated in a single process with both Kinect SDK and OpenNI[1]. Furthermore, there can only be a single depth-camera connected to a given USB controller. This limits most computers to only two Kinect devices. For this reason, a client-server architecture has been implemented to enable multiple depth-cameras to be used.

Each client can run on a separate node and transmit information to the main server. This allows each device's computer to execute the skeleton tracking algorithm and transmit its results to the main server, thus lightening the computational load of a single machine, and allowing for one machine to apply the following inverse kinematics steps without interference.

The first device to connect to the server is chosen as the main device, and it will serve as a reference for other devices. If the server has a physical device connected to it, it will choose that device as the reference. The reference device's coordinate system is then used as a base for the other devices. We gather enough joint data to estimate the transformation matrix that will allow us to pass from each device's coordinate system into the reference's coordinate system. We gather data points for each skeleton joint produced by OpenNI and feed it into Point Cloud Libraries (PCL) Rigid Transformation estimation algorithm (Documentation, 2012) to get a transformation matrix from each device to the reference device. The main device's transformation matrix is also computed to verify the validity of the resulting transformation matrices. This reference matrix needs to be within a threshold of 10e-3 of the 4x4 Identity matrix for the estimation to be considered successful. Using 1000 points of data for each joint the threshold is always reached for the reference device and the transformation is considered to be valid.

The estimated matrix will assume that the coordinate space has a linear scale, which isn't necessarily the case due to lens distortion. A study was made on the calibration parameters of the Kinect device and has shown that errors come from three main sources: the sensor itself, including lens distortion, the measurement setup and the object's surface properties (Khoshelham, 2011). This same study shows that the random error on the depth measurements follow a quadratic curve that can reach up to 4cm at the maximum range of 5m. In our application, to leave enough space for movements, we need the subject's body to fit inside the camera's field of view. A distance of 3 to 4 meters is sufficient for this, leading to a random depth measurement error of +/- 1.35 to 2.4cm. Considering that the error is a random error independently produced by each camera's depth measurement, this error represents a worst-case scenario for the depth error because the depth coordinates are averaged.

Finally, the skeleton data fed to the server is transformed into the reference coordinate space

---

[1] As of October 2012, the Kinect SDK supports multiple Kinects

and combined into a single skeleton for integration with HumanIK. This is done using a simple average of the input skeletons. Confidence levels produced by OpenNI are taken into consideration; Joint coordinates with a confidence level bellow 75% are ignored. It is necessary to note that even though a confidence threshold is applied, some "bad" values still make it through with a confidence level of 100%. The average pose calculation serves the double purpose of reducing position errors on valid data and of smoothing the resulting animation when a false positive appears, and provides a better skeletal base for which to solve the inverse kinematics equations on (see Figure 5.1).

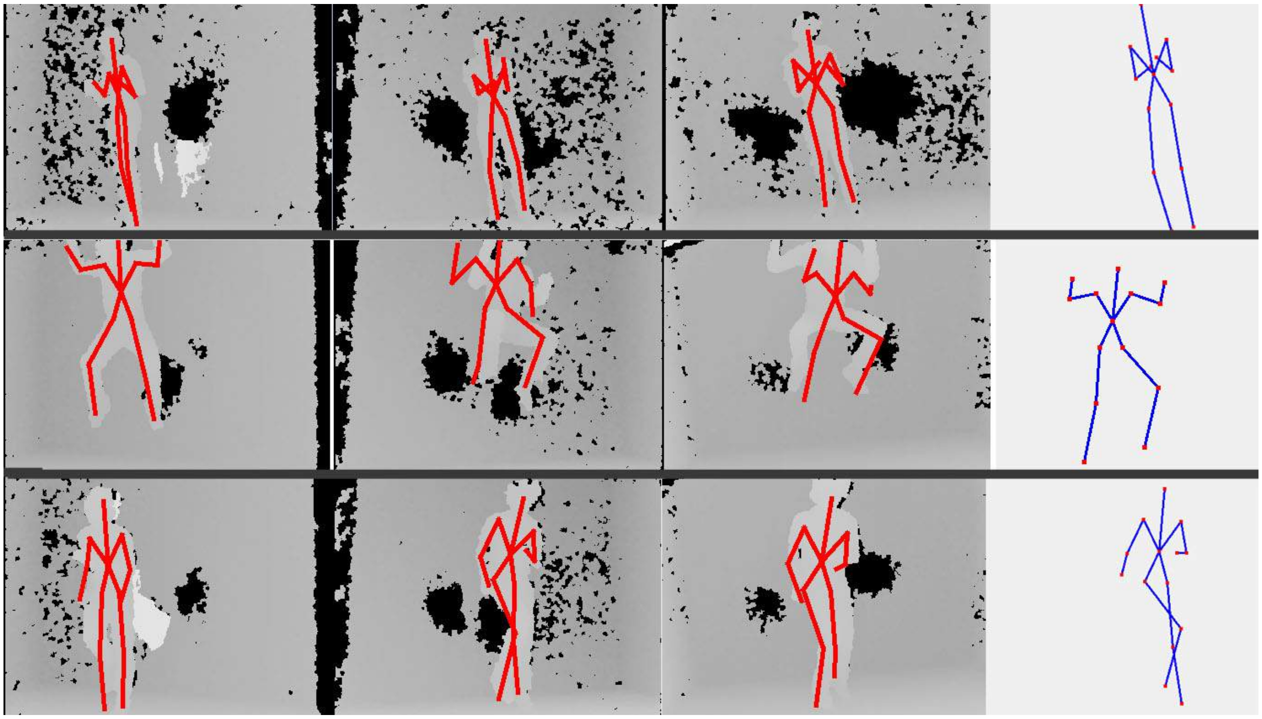

Figure 5.1 3 Kinect inputs and the resulting combined skeleton. All three examples show significant visual improvement.

Depth-cameras, including the Kinect, are active sensors, meaning they both emit and receive a signal. The Kinect camera uses a 830nm wavelength infrared laser to scan the scene in front of it. Its sensor captures the infrared light that bounces off the scene to produce a depth image.

When using multiple depth-cameras, each device will scan the scene with its infrared laser at a frequency of 30Hz. Because all devices have the same wavelength, they interfere with each other. The interference can come from both lasers combining before they hit the scene or by each Kinect capturing the emitted light of another device. In any case, the resulting noise limits the number of devices that can be used simultaneously. A study showed that a maximum number of four devices can be used while still reliably tracking a human skeleton (Schröder et al., 2011). Our testing shows that we can use a maximum of three devices at the same time, in our current setting. This limit is set by the ability of the PrimeSense library to generate a skeleton that can be used by HumanIK.

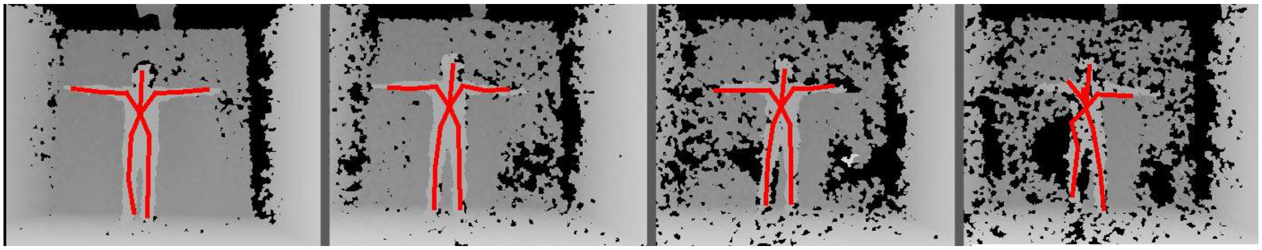

Figure 5.2 The resulting image and skeleton captured while using 1, 2, 3 and 4 Kinect devices. We see a gradual increase in interference, resulting in noise on the image and a incapability to perform the skeletal tracking.

We have extended the experiments described in Schröder et al. (2011) and found out that the environment, the subject, the setup of each device, and the texture of the surfaces can influence the results of the skeletal tracking. Another study confirms that material specularity and narrow angles lead to more errors in depth pixels (Berger et al., 2011). In our completely isolated white environment, a maximum of 3 Kinects can be used while maintaining robust skeleton tracking. Furthermore, the Kinects need to be separated by at least 1 meter from one another to effectively track the human skeleton. In addition, it was found that placing the depth-cameras at different heights increased the reliability of the skeletal tracking. With better conditions, four Kinect cameras are able to simultaneously track a human skeleton, but it is dependent on too many factors to be reliably reproduced within our staging areas, where we have limited ourselves to three devices, so that the input for the following HumanIK solving steps is stable (see Figure 5.2).

We have also found that when capturing an animation where a user jumps, the resulting vibrations reduce the noise in the captured depth images. In fact, vibrations have been found to reduce the noise in Kinect devices (Maimone and Fuchs, 2012). This is due to the fact that when vibrating, each device will see it's point pattern clearly while seeing a blurred version of the other device's patterns.

With all this information in mind, we have compared the resulting joint positions of one, two and three Kinect devices to the real joint positions, captured with a Flock Of Birds (FOB) system.

5.3 HumanIK Integration

The next step in the proposed MoCap technique is to use the depth-camera feeds to drive the human skeleton model built in Autodesk's HumanIK library. This library implements standard inverse kinematics equations and layers human constraints and limits which have been experimentally deduced in collaboration with medical experts.

HumanIK requires a character definition, characterized in a T-stance, that will serve as a base for our solved IK skeleton. The characterization is done using the subject's proportions, which requires a calibration. During this calibration, proportions are measured and a base skeleton is modified to reflect the measurements made. For the calibration, the user must be standing upright at full height. The subject's ankle-to-shoulder distance divided by the reference skeleton ankle-to-shoulder distance serves as our base transformation unit from the pre-defined skeleton to the subject's proportions. Other joint positions are then calculated according to this unit as well as other measured proportions, such as the elbow-to-wrist length. Rotations are pre-defined by the reference skeleton and are not modified by depth-camera calibration data.

Once the calibration is complete, the HumanIK pose is ready to be driven by the depth-camera combined skeleton. Each frame requires an input pose and effectors. The input pose is always the previously calculated pose. In this "Use Previous Pose" method, the effectors, which will influence the positions and rotations of each joint according to given weights and parameters, are initialized from the combined skeleton coming from the depth-cameras. The solved HumanIK skeleton is then computed on the entire skeleton. The calibration pose is chosen as the first input pose, and each new solved pose is the input pose for the next frame.

Another approach to using HumanIK in a depth-camera driven motion capture system is to completely drive the animation using the depth-cameras and use HumanIK to correct poses that are impossible for the human body to reproduce. This "Combined Skeleton Driven" approach shows much less deviations from the input skeleton, but leads to many unnatural poses within the animation.

In this approach, both the input pose and the effector positions are directly taken from the combined skeleton of the depth-cameras. The solving step will then set minor adjustments to the resulting pose so that the skeleton respects constraints imposed by HumanIK.

HumanIK also has built-in functionalities that allow human limits to be applied to a pose and blend them with the pose's relax state, all using the implemented inverse kinematics equations. Applying human limits to a pose will force joint translation and rotations of a given pose into the closest state that can be replicated by a human body. Both OpenNI and

Kinect SDK do not have anything like this built into their skeleton tracking algorithm, so the output skeletons may not represent what a human body can do.

The relax pose of a characterized skeleton is a pose where all joints are situated between their minimum and maximum extensions. Partially blending the resulting animation with the relax pose (10%/90%) adds realism and smoothness to the rendered animation, but decreases accuracy due to dampening derived from blending actual position with a static pose. Increasing the blending ratio increases the dampening in the animation and reduces the freedom of joint movements. These two constraints to the skeletal animations, “Apply Human Limits” and “Relax Pose”, prove to be an effective way to get a realistic animation without sacrificing too much accuracy.

As shown in Figure 5.3, applying “Human Limits” and “Relax Pose” constraints to input skeletons results in slightly improved poses. For most input skeletons, the resulting pose is slightly improved. For some input skeletons where the pose produced are impossible for a human, the resulting skeleton is dramatically improved.

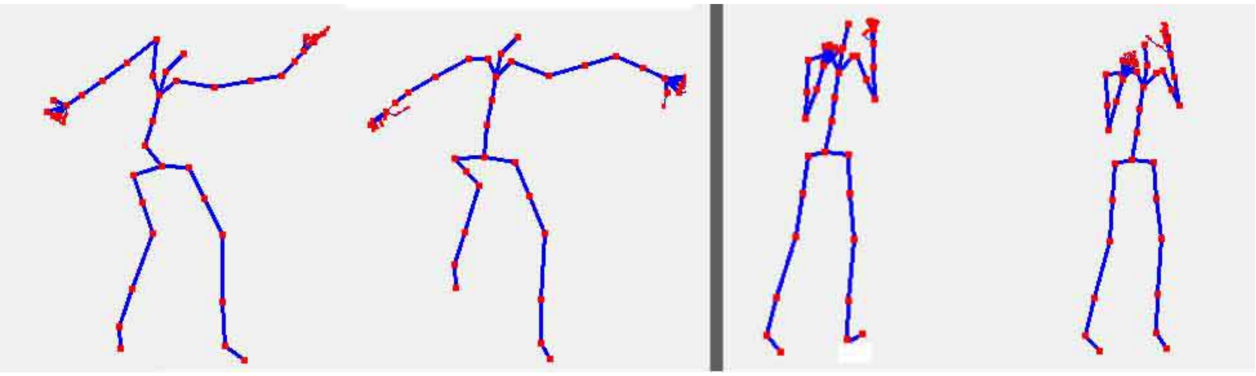

Figure 5.3 Two poses with (right) and without (left) human limits applied. On the left side, we see the shoulder has been significantly moved. On the right side, the left foot, which was originally twisted around the ankle, is corrected.

Unfortunately, these constraints also caused slight hyper-extensions to have a dramatic effect on the solved skeleton, and showed huge jumps in both positions and rotations in the animation, as well as possible but unnatural poses for a human being, such as an “A-Frame” at the knees (See Figure 5.4).

Experimental results show that using the “Apply Human Limits” and “Relax Pose” built in functionalities are very circumstantial. These two features can be safely disabled because HumanIK has built in limitations for the human body that are applied during the IK solve, using the effectors, but can be activated if some of the movements on the resulting animations result in impossible poses. The choice to use this feature will be dependent on the recorded

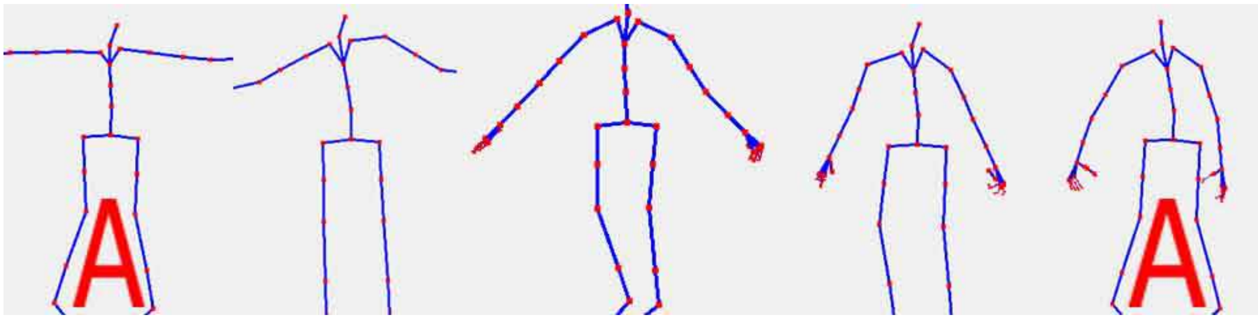

Figure 5.4 “A-Framing” and leg jitters in consecutive frames resulting from the “Apply Human Limits” and “Relax Pose” HumanIK functionalities.

movement, the resulting animation and the judgment of the user.

Although the skeleton data from OpenNI only has 15 joints, the full skeleton definition in the HumanIK library can have up to 170 joints. By defining a complex HumanIK skeleton, driven by the camera input, we can solve positions and rotations for all the extra joints that are not provided by OpenNI or the Kinect SDK. The 15 input joints, used as effectors, influence the nonexistent OpenNI joints added with HumanIK and add extra movements and information that is not captured by the depth-cameras. The extra joint positions and rotations are complete guesses made by HumanIK with the information provided by the effectors and the limit constraints set on the HumanIK joints. An instance of this effect can be seen in the spine and shoulder movements (See Figure 5.5).

In the next section, we discuss how additional data can be extracted from the depth information, deduced from other joint positions or specified by the user. This additional information can be used in the same way as the input skeleton, to drive the HumanIK solver and add realism and accuracy to the final animation.

5.4 Complementing HumanIK

The calculated poses from skeleton merging and HumanIK integration shows a clear improvement from the raw depth-camera skeleton inputs. Despite this, the resulting animation is not perceived as a convincing human being, due to stiffness in outer limbs and lack of fluidity in the movements. The lack of information for extremities renders a static animation that seems stiff, looking more like a lifeless robot than a human being. To fully take advantage of the capabilities of the HumanIK library and the human model it provides, we need to look for information in the depth images so to feed it and drive the human model in HumanIK. Extracting additional data to drive the head and hands are crucial for a realistic animation.

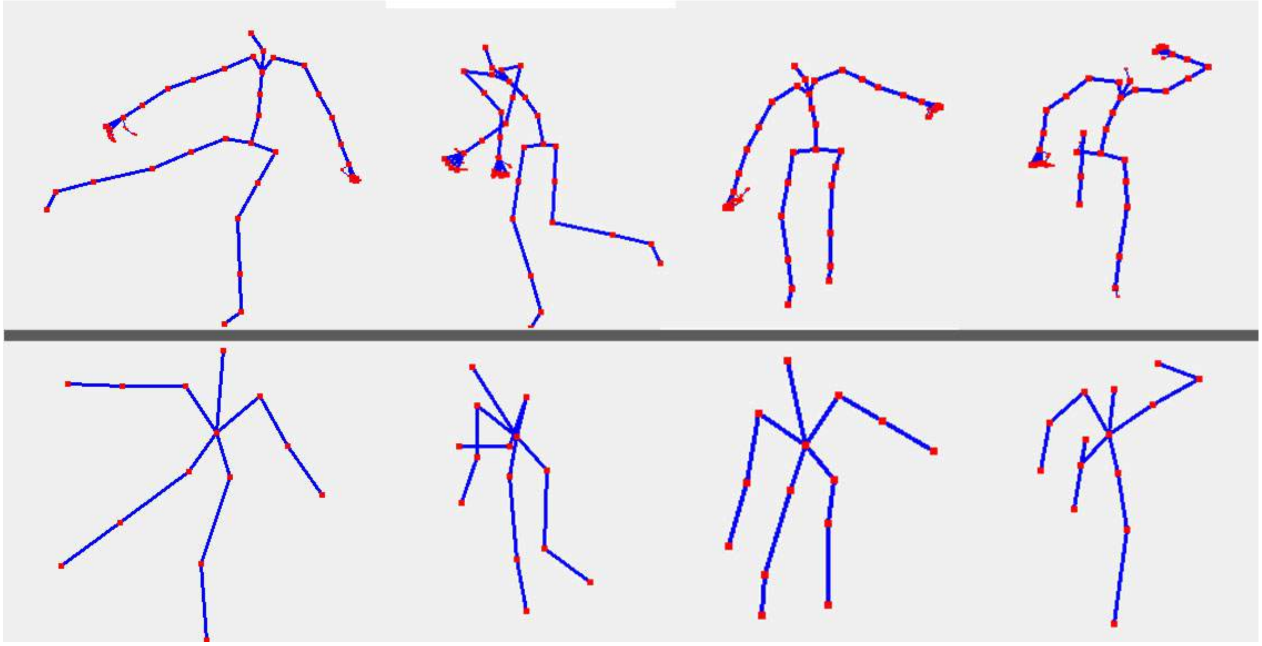

Figure 5.5 Full HumanIK Skeletons versus their respecting basic combined Kinect skeletons.

The lack of resolution in depth at the required distance to capture the entire human body prevents the capture of accurate data for extremities, but this data can be deduced from the depth images or specified by the user. We look at two specific elements which limits the realism of the resulting animations: hand states and head movements.

To remove head stiffness and add fluidity to the neck rotations, we implemented a “Look At” feature. The user can select a “LookAt” option that will determine the direction towards which the animated character will look. This takes advantage of the function HumanIK provides, reinforcing the tie of our technique with inverse kinematics. Three points have been implemented to drive the “LookAt” function: Hands, Far and Near. The head effector rotations are then set to reflect the target point where the subject would look towards during the captured animation. For example, if the subject is manipulating an object, he would look at his hands, if he is pursuing someone, he would look far away, etc. The added head movements greatly adds to the animation realism (See Figure 5.6).

The additional information to drive the hands is derived directly from the depth image provided by the cameras. Once again, we take advantage of the capabilities of HumanIK to define default constraints on joints, particularly in hands and fingers, to feed more information to the library, and provide enhanced results compared to a simple depth camera capture. The skeleton from a device is first used to locate each hand in the depth image. From the center point of each hand, concentric circles are drawn and depth data on the circumference of

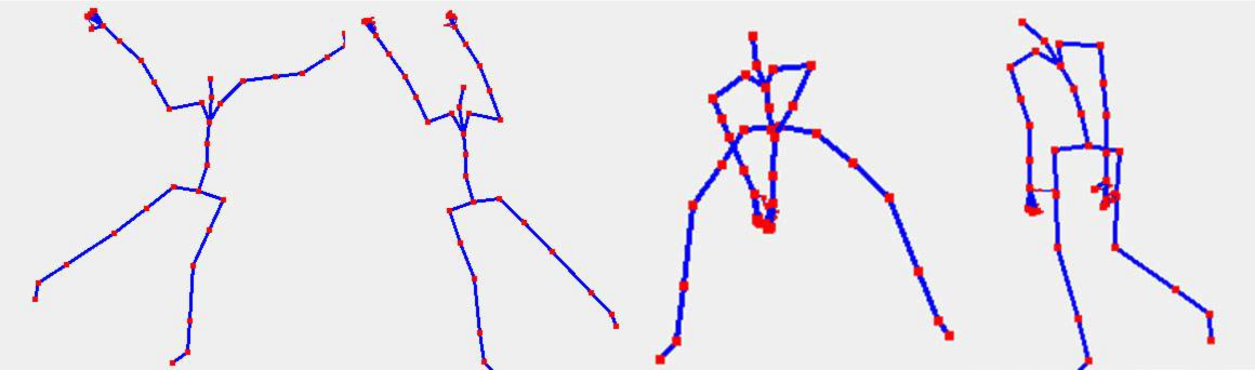

Figure 5.6 Four poses with different “LookAt” target. From left to right: No target, a point between hands, a point far away, a point close to the subject on the ground.

these circles are analyzed. First, the largest full circle is found. A full circle is defined as a circle where all the depth points on its circumference are approximately at the same depth as the hand itself. The next circle is then analyzed, and three cases can happen; If the points on the circumference of the circle that are at hand depth reach a threshold, the hand is opened; If the points on the circle circumference that are not at hand depth reach a second threshold, the hand is in a fist; In the case where the point count does not reach any of the two thresholds, the previously found state is assigned.

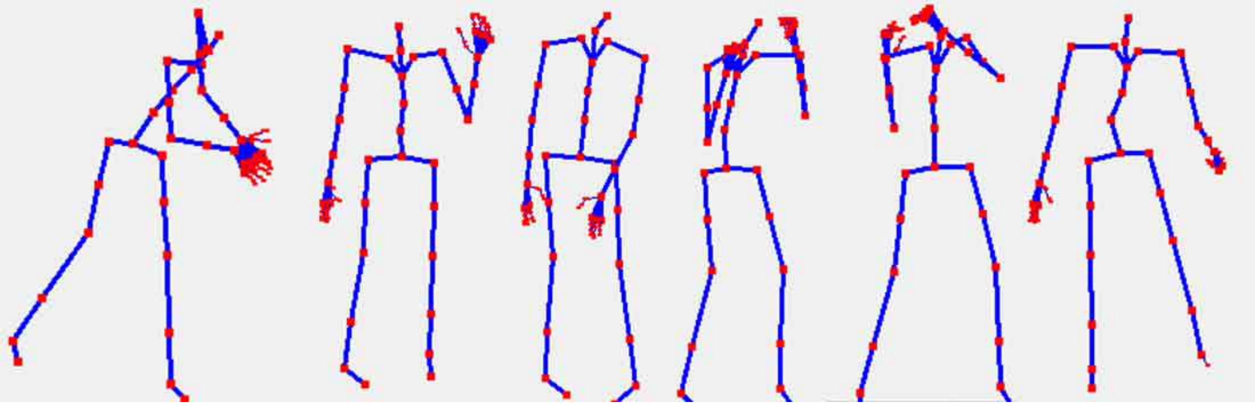

Figure 5.7 A few resulting hand positions and states.

This method, although not very precise for a single frame, averages a better result than a random hand state. If the hand state is guessed at random, its accuracy would be approximately 50%. We captured sequences where the hands remain closed and remain opened, and we found that hand state was accurate 66.81% of the time, or a 16.81% improvement to random guesses, despite the unstable nature shown for a single frame. The HumanIK finger effectors are then set either at the extremities of the fingers or at the center of the palm of the hand, creating a closed or open hand when the pose is solved.

The additional data gathered by these two simple techniques, fed into the inverse kinematics and human model provided by HumanIK, have proven to add a natural feeling to the resulting animations (See Figure 5.7).

5.5 Filtering and Smoothing

As we mentioned in the previous section, the calculated poses from skeleton merging and HumanIK integration shows a clear improvement from the raw depth-camera skeleton inputs. Despite this, the resulting animation shows spikes in movements and a lot of jittering, such as big jumps in positions between frames which are not possible for a human. The shortcomings of inverse kinematics are felt here; It does not provide any temporal restraints, so that movement limits and smoothing has to be implemented to complement this lack.

The first step taken to remove spikes is to apply the default smoothing provided by OpenNI and KinectSDK, which is based on the Holt double exponential smoothing method (Holt, 2004; Winters, 1960). Although this does give a significantly smoother animation and almost completely removes movement spikes, it greatly reduces movement resolution and accuracy. In motion captures where the subject is running, even with a relatively small smoothing factor, the nature of the movement is lost to the average viewer (See Figure 5.8).

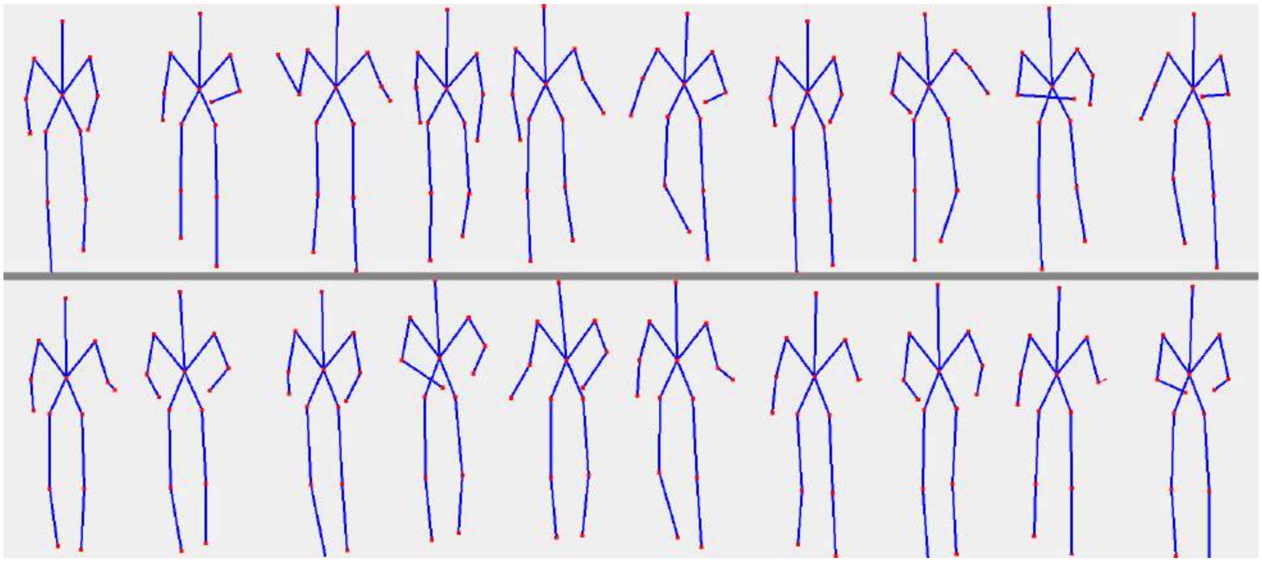

Figure 5.8 10 random frames of a run animation with(bottom) and without(top) Holt smoothing on a combined skeleton. Although the third frame of the top sequence shows an invalid “spiked” frame, the animation is easily identifiable as a running sequence. The same cannot be said for the bottom sequence.

A more successful approach to reducing movement spikes within an animation without sacrificing too much accuracy is to limit a joint's displacement between frames. By defining the skeleton hierarchy, the norm of each joint's displacement is calculated from frame to frame. The displacement is compared to a threshold value, defined for each individual joint, which limits a joint's displacement relative to its parent joint.

Filtering animation spikes can be done before or after solving our HumanIK pose, and both situations have been implemented. We hypothesized that smoothing out animations before the solved HumanIK skeleton would transfer into smoother resulting animations and would add accuracy compared to a smoothing done after the HumanIK solving step. The results were similar to the Holt smoothing method on the skeleton; the resulting animation was smoother but accuracy was lost.

This is also applied to finger movements. The unstable nature of the hand state algorithm for a single frame produces immense jumps in finger positions between states. By applying this movement limitation to all finger joints, these big jumps are completely removed and the alternation between hand states amplifies the smoothness and natural feeling of a human hand by adding movement.

Regardless of when filtering is applied, one particular case of jittering is seen in the elbows and knees of the rendered skeleton. This is due to the fact that a slight change in positions can trigger a big variation in rotation between frames. For example, if an arm is fully extended, the shoulder, elbow and wrist will follow a straight line. If the elbow is then slightly over extended (bent the other way, as a double-jointed individual could do), the HumanIK solver will automatically rotate the entire arm the other way to match the new elbow position. This behavior was apparent when the HumanIK "Apply Human Limits" functionality was activated. To counteract this jittering in the elbows and knees, those particular effectors were given less freedom to move and rotate, as well as given less influence over their parent and children nodes.

As noticed above, most jittering happens when joints are hyper-extended, when the skeleton tries to reach too far, or when bones try to rotate too much. For this reason, bone lengths have been limited to their calibration length. Since HumanIK already limits bone lengths, this limitation is implemented before the solving of the new pose with HumanIK, directly on the combined Kinect skeletons. The results show a generally smoother animation, but precision is once again lost.

5.6 Results

The motion capture technique described has many independent elements which, once combined, produce a realistic and eye pleasing animation sequence. Each addition to the system has been independently tested or measured to insure that they individually add quality or precision to the rendered animation and that they don't interfere with each other.

To test accuracy, short motion capture sequences were taken simultaneously with our technique using multiple Kinects, and with a Flock Of Birds (FOB) system by Ascension. The latter has a positional resolution of 0.8mm, an accuracy of 2.5mm and an update rate of 100Hz. This system will serve as our benchmark to measure accuracy. It is far more accurate than the Kinect which has a depth measurement error of 24mm at the required depth of 2 meters for our capture. For each capture, the positions of the right or left wrist was measured, in reference to the shoulder and the chest positions.

The measurements done with the Kinect and the FOB system output curves which indicate the position of the sensor or joint over time. These curves are then superimposed and quantitatively compared to one another. There are a few limitations to comparing values for each curve. The first is that the FOB system and the Kinect do not output at the same frequency, meaning we cannot compare a position at the exact same time. Since the FOB has a frequency far higher than the Kinect, for each joint position given by the Kinect, we compare it to the closest time value given by the FOB sensor. Another limitation is that the output rate for the Kinect is not constant. In fact, when comparing the graphs for an up-and-down wrist movement in the axis pointing up, we found that the waves described by the Kinect curves shift in wavelength drastically more than the ones produced by the FOB system. This can be caused by a number of factors such as an unsteady frame rate or the middleware computing time.

Because of this, we have taken into account two sets of measurements when comparing graphs, shown in Figure 5.9. The first set measures absolute position difference for a given Kinect time value and the closest matching FOB value. The second set measures the shortest distance between a given point in the Kinect and the FOB curves.

The two measures show their own advantages and limits (See Figure 5.10). The first of the two, the absolute difference, gives the upper limit on the accuracy of our measurements. It also shows the accuracy of the animation curves through time. However, it's very sensitive to spikes and deviations between the curves, which will rapidly increase in value. A lower measured value always means better accuracy during the capture.

The second measurement, the closest point difference, gives a lower limit on the accuracy of

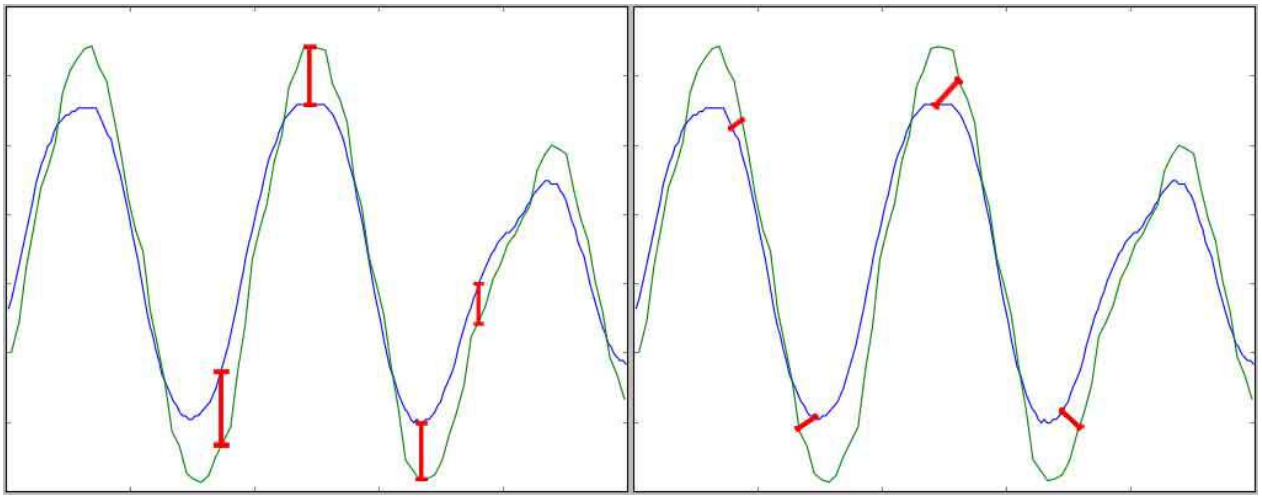

Figure 5.9 A curve of wrist movements in the Y axis (height) showing the two measurements we take when comparing graphs. Left: Absolute Difference between matching times. Right: Closest Point between graphs.

the measurements. It is not as sensitive to spikes and deviations and somewhat ignores time distortions. A lower value on this measurement does not necessarily mean better accuracy on the captures, but means that the two curves look more alike.

With these two measurements, we will be able to get an upper and lower bound on the animation accuracy, as well as objectively analyzable data for a single curve. It is also worth noting that accuracy values for both measurements are very sensitive to the captured clip and can drastically change between different clips, which is why the same clip is used to measure each independent element in our motion capture method. In the case where the same clip cannot be used, we increase the number of clips.

Finally, for each curve we count the number of spikes and deviations. Spikes are defined as a variation of more than 10cm over 0.13 seconds (4 frames on the 30frames per second Kinect camera) will be counted. This means that each local maximum in animation curves will usually show up as a spike as well as imperfections in a curve. The difference in the number of spikes will be the significant value for spike measurements, and it is defined as the number of spikes in the Kinect measurements minus the number of spikes in the FOB measurements. A deviation between curves is when one curve drastically differs in appearance to the other. For example, if one curve shows a “mountain” and the other shows a “valley”, this counts as a deviation.

The subjective aspect of animation quality has been tested with a series of blind tests. A group of 66 students following a bachelor’s level computer graphics class volunteered to watch

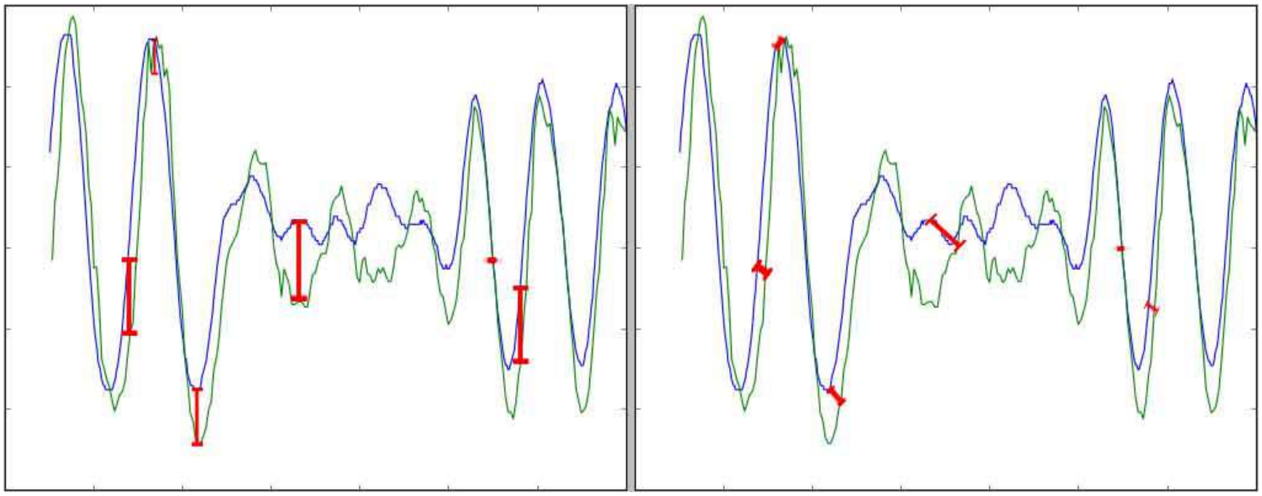

Figure 5.10 A curve of wrist movements in the Z axis (depth), with deviations and spikes, showing the limits of the two measurement methods. Left: Absolute Difference, showing sensitivity to small time differences, spikes and curve deviations. Right: Closest Point, showing very small values for deviations and time differences.

a series of animations, two at a time. In each comparison test, a feature was turned on or off and each student was asked to chose their preferred animation, as well as comment on the animations.

5.6.1 Accuracy of using multiple Kinect depth-cameras

The first element that was tested was the accuracy of the Kinect depth-camera, and the accuracy of multiple cameras. The accuracy measurements were done three times, on separate clips. For each of these clips, only the combined skeleton was considered. The HumanIK solving step was disabled and no additional processing was done. The results are displayed in Table 5.1.

Table 5.1 Accuracy comparison for captures with 1, 2, and 3 devices

# of Devices	1	2	3
Absolute Difference (cm)	14.4	13.2	15.7
Shortest Distance (cm)	6.9	4.4	5.1
Spike Difference	7	2	11
Deviations	3	2	3

From the measured data, we can see a clear improvement between the 1-device and the 2-device captures, but also a big deterioration when we capture with 3 devices. The deterioration from 2 to 3 devices is due to the amount of noise in the depth images and the inability to accurately capture the skeleton, as seen in Figure 5.2. Perhaps that better capture conditions and a less reflective and less uniform environment could improve the 3-device capture measurements, as previously discussed, but our current setup led to bigger error measurements than the single device capture. However, we see a great increase in accuracy when going from 1 to 2 devices, as well as a reduction in spikes and deviations. Despite the interference, the skeleton detection with 2 device is reliable and the combination of two skeletons compensates when one device fails to recognize parts of the skeleton.

Note that the comparison between a single device and multiple devices was not submitted to a blind test. This is due to the fact that it is impossible to capture the exact same sequence with one, two and three devices separately, as there will always be differences in the subject's movements. In addition to this, if we were to capture an animation with two devices and remove one of the inputs during playback, the single-device sequence would still have the noise from the two-device capture in it, making the comparison impossible.

5.6.2 Comparison with HumanIK integration

The next element to be tested was the comparison between an animation driven by the previously calculated pose and updated with HumanIK versus an animation driven by the combined Kinect skeletons and corrected by HumanIK, as described in Section 5.3.

Table 5.2 Accuracy comparison for HumanIK integration methods

HumanIK integration	No HumanIK	Use Previous Pose	Combined Skeleton Driven
Absolute Difference (cm)	14.2	12.6	11.2
Shortest Distance (cm)	6.5	4.3	3.3
Spike Difference	7	9	3
Deviations	3	2	2

The same clip was used to compare the two methods. We can see that both methods show an improvement on accuracy compared to the Kinect only (see Table 5.2). This is interesting considering that integrating HumanIK does not add any positional data to the animation, but simply follows a human model. We can also see that one curve deviation was removed from the original clip. We also observe that the "Combined Skeleton Driven" approach shows better accuracy and reduced the spike difference between clips.

For this blind test, volunteers were shown the same animation, captured with a single device, where the only difference was the way in which the animation is driven. The first technique, where the previous pose is used, obtained 50 votes, compared to 16 votes where the combined skeletons drove the animation. Volunteers who picked the second method commented that the first seemed unstable, and those who picked the first shared the general opinion that it looked more active.

These accuracy and blind test results are opposite to each other, meaning that the choice between driving the HumanIK model with the previous pose versus the combined skeleton sacrifices accuracy for visual quality and smoothness.

5.6.3 Measuring HumanIK functionalities effects

The third aspect to be tested was the “Apply Human Limits” and “Relax Pose” built-in functionalities of HumanIK. First, we measured differences in joint position for each frame, with and without these functionalities activated. We observed that 55.20% of frames (3437 on 6227) have at least one joint moved by at least 10cm from their original positions when the features were enabled (see Table 5.3). The analyzed frames were taken from 5 different captures, with varying movements and Kinect configurations.

Table 5.3 Accuracy comparison for the “Apply Human Limits” and “Relax Pose” features

Apply Human Limits	Disabled	Enabled
Absolute Difference (cm)	12.6	10.3
Shortest Distance (cm)	4.3	3.8
Spike Difference	9	6
Deviations	2	2

Once again, the same clip was used to compare the two methods, and we see an increase in accuracy and a reduction in spikes, despite HumanIK not having any additional information to render each pose.

Finally, a blind test has been administered to compare animations with and without the features. Three different sequences, each with and without the functionalities activated, were shown to the volunteers. These sequences were captured with a single device, and feature the cases previously discussed: A-framing, impossible rotations, and major position corrections. The results are divided: 52 votes for the features enabled, and 48 votes for the features disabled. Some volunteers explicitly pointed out the flaws discussed for each sequence, while others were not bothered by them, but pointed out that the overall animation looked more realistic. Most comments against this feature mentioned instabilities and jitters in the animations.

5.6.4 Bone length limitation

Next, the limitation of bone lengths before the HumanIK solving step was tested.

Table 5.4 Accuracy comparison for bone length limitation

Bone Length Limitation	Disabled	Enabled
Absolute Difference (cm)	12.6	12.4
Shortest Distance (cm)	4.3	4.6
Spike Difference	9	8
Deviations	2	2

We saw no significant increase in accuracy when we enabled this feature (see Table 5.4). The blind test used the same clip with the feature enabled and disabled. The opinions were split: 28 volunteers preferred the clip with the feature disabled and 38 preferred with the feature enabled. When asked why, the volunteers said that with the feature disabled, the animation looked more lively and that the other clip seemed smoother.

5.6.5 Movement limitations

The next aspect to be analyzed is the joint displacement limitation, and where it should be applied within the animation process (See Section 5.4). Two possibilities were considered: before and after the HumanIK solving.

Table 5.5 Accuracy comparison for animation smoothing

Smoothing	None	Before	After
Absolute Difference (cm)	12.6	14.0	13.5
Shortest Distance (cm)	4.3	4.9	5.5
Spike Difference	9	5	7
Deviations	2	2	2

The accuracy results show that smoothing the animation reduces the accuracy of the capture (see Table 5.5). This is due to the fact that the skeleton cannot always reach the captured position at each frame due to the movement limitation in the smoothing method.

The blind tests featured the same animation, captured with a single device, with the smoothing enabled and disabled, as well as smoothing done before and after the HumanIK solving stage. The clip with smoothing enabled received 51 votes, compared to the clip without smoothing, which got 15 votes. In the next test, the results are 18 votes for smoothing before the HumanIK solving and 48 votes for smoothing after. Most volunteers shared the opinion that pre-smoothing seemed to dampen the animation and that post-smoothing animation looked livelier.

Once again, we see a sacrifice between accuracy versus visual quality in the rendered animation.

5.6.6 Complimentary data

Next, hand state accuracy has been measured by recording two distinct sequences; one where the subject keeps his hands opened and the other where he keeps them closed (See Section 5.4). For these two recordings, hand states reached an accuracy of 66.81% (1512 on 2234 frames), where 41.22% of frames found undetermined states and relied on the previously found state. These sequences were recorded with a three depth-cameras to simulate worst-case scenario noise, but the hand state recognition was done only on the main depth sensor. Although it has not been tested, it is possible that the accuracy could be increased by implementing the algorithm on the client side and sending the result for each device to the server.

A blind test has been conducted with the “LookAt” and hand state features enabled and disabled (See Section 5.4). The two sequences, captured with a single device, included filtering after the HumanIK solving step to smooth out finger movements. The results were that 64 out of 66 volunteers preferred the animation with the features on. In addition to this, comments hinted to better head and finger movements (or presence of) and not one volunteer mentioned that the hand states did not seem accurate. The two volunteers that preferred the animations with the features disabled said that the fingers moved too much.

5.6.7 Final results

Each component has been analyzed individually as to integrate the ones that show an improvement compared to simple depth-camera capture into our proposed solution. We have concluded that 2 depth-cameras have given best results. We have also shown that using the previously calculated pose as a seed to the inverse kinematics is the best approach, since it shows both greater accuracy and a better visual quality compared to using the current pose. We also activate the human limits, keep the additional data found for head movements and hands, limit bone length and activate our movement filter.

We observe significant improvements: every measured aspect shows improvements compared to normal capture (see Table 5.6). The upper limit on position error is improved by 30.9% and the lower limit by 47.7%. Both the number of spikes and the deviations from the “Flock

Table 5.6 Comparison between simple capture with a single depth and our proposed technique, with all components showing improvements activated

	Single Kinect and Normal capture	Proposed technique with 2 depth cameras
Absolute Difference (cm)	14.2	9.8
Shortest Distance (cm)	6.5	3.1
Spike Difference	7	4
Deviations	3	2

Of Birds" capture have been reduced. Furthermore, the proposed technique with all active components has been submitted to the blind test. All 66 volunteers have chosen the proposed technique as their preferred capture.

5.6.8 Using the results

Finally, our resulting animation is a series of frames containing position and rotation data for every joint available. This data can be exported in various formats, and interpreted by professional motion capture software such as Autodesk's *Motion Builder*.

The chosen export format is FBX, which can specify an interpolation method when reading the data. For our purposes, we specify a cubic interpolation, which will smooth out the animation between frames, resulting in a cleaner animation when viewing it. In fact, a cubic interpolation between frame shows a dramatic improvement in the smoothness and overall quality of the animation. In a blind test comparing no interpolation versus cubic interpolation, all 66 volunteers chose the latter. The resulting file can then be read by *Motion Builder* and can then be used to provide extra smoothing and filtering as well as mapping the animation to a skinned mesh for viewing, like any other motion capture file.

5.7 Conclusion

The proposed motion capture technique using multiple depth-cameras and Autodesk's HumanIK library has shown promising results. We have found that using multiple Kinect cameras requires communications between processes and computers, which can be handled with a client-server architecture. We have also shown the limitations of working with multiple depth-cameras, that some usage techniques and setups can increase accuracy and improve the depth image produced, and that increasing the number of devices results in better accuracy, as long as the skeleton detection is reliable.

We have tried multiple ways to integrate Autodesk's Human Inverse Kinematics library with our technique, and we have found that HumanIK will increase both accuracy and quality of rendered animations. We have also found that some elements of the HumanIK integration are a trade-off between spacial accuracy and animation quality (smoothness) of the resulting animation, such as using the previous pose to drive the HumanIK model or using the resulting combined skeleton.

We have found that additional data can be found or inserted into the HumanIK model to help drive parts of the skeleton like the hands and head, which result in more realistic animations. Finally, we also found that smoothing the animations generated by depth-cameras is necessary due to faulty skeleton detection, false joint positions and jittering.

We have only touched the surface of the potential of depth-cameras in motion capture. As camera resolutions and quality increase, we will see increasingly better results. Furthermore, our technique could potentially be improved in several ways.

First, a better joint selection process could be implemented when combining skeletons. Even if OpenNI's confidence level on a joint is of 100%, if the joint position from one device is largely different from the other two devices, this joint could be discarded.

Another improvement would be to place the depth-cameras all around the user and extract a more accurate skeleton from the resulting point cloud. This has the potential of increasing the accuracy as well as the confidence levels of joints.

The capture setup could also be improved. Our test environment featured white walls, which would reflect the infrared wave emitted by the Kinect devices, resulting in additional noise. Furthermore, we have seen that adding vibrations to the devices could improve the resulting depth images. With a better setup, it is possible that more devices could be taken advantage of, without drowning in interference noise.

These two examples show that there is still room for improvement. Nonetheless, we have shown that combining multiple depth-cameras with the Autodesk's Human Inverse Kinematics library model can produce some significantly more pleasing animations.

Acknowledgment

Autodesk for financing, office space, tools and support.

CHAPITRE 6 DISCUSSION GÉNÉRALE

6.1 Introduction

L'article présenté montre et décrit les étapes suivies afin d'implémenter la technique de capture proposée. On y décrit également les résultats obtenus et les limites rencontrés. Nous allons maintenant élaborer sur les résultats obtenus et en tirer des conclusions sur leur utilité et pertinence.

6.2 Résultats

La combinaison des caméras 3D et de la bibliothèque HumanIK d'Autodesk a donné des résultats très diversifiés. Autant du côté quantitatif que qualitatif, certaines méthodes améliorent les captures de mouvements, certaines nuisent au but recherché, mais toutes sont sujettes à questionnements et discussions. Nous observons en effet différentes conséquences à l'ajout des contraintes du corps humain à la capture de mouvements à partir de caméras 3D. La présente section élabore sur les résultats obtenus.

6.2.1 Précision de la capture

La précision de la capture de mouvement a été mesurée en comparant la capture effectuée avec la méthode proposée à une capture effectuée avec un système *Flock Of Birds* (FOB) par Ascension. La même séquence de mouvement a été capturée simultanément et les résultats ont été compilés afin de permettre leur analyse.

Chaque élément de la technique de capture a été évalué indépendamment pour mesurer leurs effets individuels sur le résultat final de la capture et mesurer leurs effets directs sur la séquence de mouvements résultants. De plus, toutes les composantes ayant montré une amélioration de la précision dans la capture finale ont été combinées afin de vérifier qu'elles synergisent et montrent une amélioration globale. Cette vérification finale de la combinaison des méthodes est cruciale afin de s'assurer que les différentes expériences ne se nuisent pas entre elles, mais aussi pour vérifier quelles méthodes chevauchent entre elles.

Nous utiliserons les 4 métriques expliquées dans la section méthodologie afin d'évaluer les résultats objectivement. Ces métriques sont :

1. la différence absolue entre les courbes de référence et de capture ;
2. la différence entre la plus courte distance entre les courbes ;

3. la différence entre le nombre de pics entre les courbes ;
4. le nombre de déviations entre les courbes.

Les comparaisons des méthodes se font à partir des graphiques générés en comparant les données recueillies simultanément par le système de capture Flock of Birds et les caméras Kinects.

La figure 5.9 montre un exemple d'un tel graphique. Ceux-ci sont générés en superposant la position d'une articulation sur un axe dans le temps. Par exemple, la figure 3.1 représente la position du poignet du sujet dans l'axe de la profondeur par rapport à la caméra.

Les données de positionnement pour chaque articulation ont ainsi été recueillies pour le système Flock of Birds et les caméras Kinects dans des fichiers textes et ont été horodatées selon les horloges locales des ordinateurs. Puisque les captures sont très courtes, soit de l'ordre d'une minute, nous ignorons les décalages des temps enregistrés puisque ceux-ci sont négligeables par rapport à l'erreur des caméras 3D.

Nous avons ainsi effectué 48 captures différentes, soit 4 séquences de mouvements, répétés 4 fois, avec une, deux et trois caméras de profondeur. Les graphiques ont étés générés avec des scripts écrit en Python qui superposent les courbes obtenues avec le système Flock of Birds et la solution proposée. Par exemple, dans la figure 3.1, la courbe bleu représente celle obtenue avec le système Flock of Birds et la verte celle avec les Kinects.

Les données obtenues par les caméras Kinect montrent un changement d'échelle dans le temps. Par exemple, la courbe de la Kinect est à l'avance sur la courbe du système Flock of Birds, mais est en arrière à la fin de la capture. La source du décalage peut se retrouver à plusieurs endroits. Une accumulation de délai dans le temps de traitement ou dans l'envoi des données au serveur peut en être la source, ou même une erreur dans l'horodatage même des caméra Kinects. La source du décalage n'a pas été exploré, mais plutôt différentes métriques ont étés calculés afin d'obtenir des bornes sur l'erreur du positionnement des articulations.

Nous analysons donc les graphiques pour extraire les quatre métriques. La première est obtenue en effectuant la moyenne sur les valeurs absolues de la différence entre les deux courbes à un temps donné. Sur le graphique, cette mesure est représentée par les distances verticales entre les deux courbes pour un certain point dans le temps, et fournit une borne supérieure à l'erreur de position. Encore une fois, des scripts Python ont été écrits pour faire les calculs nécessaire.

De la même façon, la deuxième mesure a été extraite. Celle-ci représente la moyenne de la plus petite distance entre les courbes. Donc, pour chaque valeur sur la courbe verte, nous trouvons la valeur la plus proche sur la courbe bleue, indépendamment du temps. Cette mesure représente donc la borne inférieure sur l'erreur de position.

La troisième métrique tente de mesurer la qualité visuelle de l'animation produite. Celle-ci mesure la différence entre le nombre de pics retrouvés sur les deux courges. Un pic est définit comme une différence de position de plus de 10 cm en 3 trames consécutives. Le nombre de pics est ainsi extrait automatiquement pour chaque courbe du graphique et leur différence est enregistrée.

Finalement, la quatrième métrique a pour but de quantifier le nombre de déviations entre les deux courbes. Cette métrique a été extraire manuellement, en observant les graphiques et en comptant le nombre de fois ou il y a une différence de concavité entre les deux courbe.

De plus, les tests à l'aveuglette ont été faits avec les deux techniques. Afin de vérifier la composante subjective de la capture de mouvement, c'est-à-dire la qualité visuelle des captures, des individus ont été demandés de donner leur opinion sur différentes séquences de mouvements capturés. Les individus sélectionnés pour prendre part dans le test à l'aveuglette sont des étudiants en génie informatique et génie logiciel de l'École Polytechnique de Montréal, suivant le cours d'infographie. Ceux-ci sont donc familiers avec les principes de base des techniques de rendu 3D par ordinateurs, mais ne sont pas des experts en animation. Le groupe était composé de 66 individus, à qui ont été présentées les séquences de capture de mouvements avec différents paramètres. Les étudiants devaient ensuite donner leurs opinions sur laquelle des deux animations présentées ils préféraient et élaborer sur les raisons derrière leur choix.

Les tests ont été administrés en choisissant au hasard des volontaires en groupe de 5. Ces 5 étudiants ont été présenté un vidéo montrant deux animations côte-à-côte. L'animation montre une séquence de mouvement élaboré contenant plusieurs mouvements tels qu'un saut, coup de point, course sur place, etc. Les volontaires regardent le vidéo jusqu'à ce qu'ils fassent un choix et écrivent leur choix sur un morceau de papier. L'administrateur du test connait en tout temps quels captures et quels composantes sont testés, mais se tient séparé des volontaires lorsque ceux-ci regardent les vidéos.

Les résultats sont enfin compilés dans le tableau 6.1

La première étape dans le système de capture suggéré consiste à évaluer le nombre maximal de caméras 3D pouvant être utilisées simultanément. L'étude mentionnée dans la revue de littérature Berger et al. (2011) démontre que l'utilisation de 4 caméras en simultanée est possible, mais que les résultats varient considérablement en fonction de l'environnement et de la texture des vêtements du sujet. Lorsque nous essayons de reproduire ces résultats, nous observons une très grande croissance dans l'interférence entre les caméras, résultant en une

Tableau 6.1 Comparaison des différentes composantes intégrées dans la solution

Composante	Différence Absolue (cm)	Plus petite distance (cm)	Différence de pics	Déviations	Test à l'aveuglette (votes pour, sur 66)	Retenue
Référence (une caméra, aucun ajustement)	14.4	6.9	7	3	0 (vs solution finale)	-
Deux caméras, aucun ajustement	13.3	4.4	2	2	-	oui
Trois caméras, aucun ajustement	15.7	5.1	11	3	-	non
Pose précédente utilisé comme source pour la cinématique inverse	12.6	4.3	9	2	50	oui
Pose courante utilisé comme source pour la cinématique inverse	11.2	3.3	3	2	16	non
Limites humaines appliqués au squelette	10.3	3.8	6	2	34	oui
Limitation de la longueur des os	12.4	4.6	8	2	38	oui
Limitation de la vitesse de déplacement avant le calcul de la pose	14.0	4.9	5	2	18	oui
Limitation de la vitesse de déplacement après le calcul de la pose	13.5	5.5	7	2	48	non
Combinaison des composantes	9.8	3.1	4	2	66	-

augmentation de bruit dans l'image, et une incapacité de reconnaître le squelette. Dans le laboratoire utilisé, nous atteignons la limite de la détection du squelette avec trois caméras Kinect. Nous avons tenter de reproduire les résultats décris dans l'étude Berger et al. (2011) sans succès. Il est possible que certains matériaux utilisés dans la recherche, les conditions d'illumination ou même la nature de l'arrière plan contribuent à la qualité de l'image de profondeur détectée.

Même si l'algorithme de OpenNI réussit parfois à reconnaître le squelette du sujet, celui-ci n'est pas suffisamment stable pour être utilisé pour la capture de mouvements, tel que montré par la figure 5.2. Nous utiliserons donc trois caméras 3D comme étant la limite supérieure au nombre de caméras 3D à utiliser pour poursuivre l'analyse des résultats.

Avec cette limite déterminée, nous pouvons procéder à l'analyse des composantes établies dans la méthodologie. Le tableau 6.1 montre une compilation des résultats obtenus, mettant de l'avant les quatre métriques mesurées, le nombre de sujets ayant votés pour la composante dans le test à l'aveuglette et si la composante est retenue pour la solution finale.

Afin d'analyser les composantes, nous avons utilisé une multitude de séquences triviales où les mouvements sont très variés. Ces séquences incluent une combinaison de mouvements simples et mouvements complexes. Par exemple, une séquence demande au sujet de demeurer fixe et de simplement bouger un bras horizontalement et verticalement en succession. Un autre mouvement demande au sujet de courir sur place, sauter, donner des coups de poing, etc. Une grande diversité de mouvements est nécessaire afin de montrer que l'amélioration, ou la détérioration, des séquences capturées sont valides dans plusieurs cas très différents.

Certaines valeurs ressortent du tableau. Premièrement, nous observons que l'utilisation de trois caméras Kinects résulte en une perte de précision par rapport à deux caméras, à cause de l'augmentation du bruit dans les images de profondeur. Aussi, à cause de cette interférence, nous ne pouvons pas utiliser exactement la même séquence de mouvement pour comparer les résultats. En effet, si on utilise l'enregistrement de deux des trois caméras, nous allons tout de même avoir l'interférence reliée à trois caméras dans les images capturées. Pour cette raison, le même mouvement a été répété 3 fois, devant une, deux et trois caméras 3D, et la moyenne des résultats des trois séquences a été utilisée aux fins de comparaison. La répétition des mouvements, leurs similarités et la constance de l'environnement de capture peuvent alors être comptées pour contrer les erreurs dues aux différences dans les mouvements du sujet.

Dans le tableau 6.1, nous observons également que des tests à l'aveuglette n'ont pas été effectués sur la comparaison entre l'utilisation de multiples caméras. Premièrement, l'amélioration entre une et deux caméra et la détérioration entre deux et trois caméra est évidente et suffisante pour justifier ne pas procéder avec les tests. De plus, la capture simultanée d'un même séquence est impossible. Puisque les caméras causent de l'interférence entre elles, la séquence de mouvement doit être capturée de façon indépendant pour une, deux et trois caméras. Le test a l'aveuglette ne présenterais donc pas la même capture et les résultats pourraient être une conséquence dans de petites différences dans les mouvements de l'acteur plutôt qu'une meilleure performance. C'est pour ces deux raisons que les comparaisons entre le nombre de caméras n'ont pas été soumises aux tests à l'aveuglette.

Il a été montré que l'intégration de la cinématique inverse utilisant la pose courante comme référence plutôt que la pose précédente donne de meilleurs résultats quant à la précision de la capture. Malgré ceci, nous retenons l'utilisation de la pose précédente comme source pour la bibliothèque HumanIK dans la solution finale, puisque les tests à l'aveuglette contredisent les résultats de précision. En effet, la méthode d'intégration utilisant la trame courante, malgré qu'elle montre une plus grande précision dans la capture de mouvement, cause certains artéfacts dans l'animation qui nuisent à la fluidité de la capture de mouvement. Ce sont ces artéfacts de mouvements qui, selon les individus interrogés, ont fait basculer les opinions. De ce fait, la majorité des sujets préfère l'animation produite avec la technique qui utilise la pose précédente comme source. Celle-ci améliore la qualité visuelle de l'animation au coût de la précision, en amenant une certaine fluidité et un aspect plus naturel à la capture finale. Puisque les deux méthodes ont montré une amélioration par rapport à la capture sans cinématique inverse, c'est donc la méthode utilisant la pose précédente qui est dans la solution finale, puisque sa qualité visuelle plaît davantage au public.

Dans les tests à l'aveuglette qui concernent l'application des limites humaines au squelette et la limitation de la longueur des os, nous observons une divergence dans l'opinion des sujets. Environ la moitié d'entre eux ont préférés les composantes activés, tandis que les autres ont préférés les composantes désactivés. Les commentaires recueillis indiquent que les composantes causent des poses maladroites dans les animations produites 5.4 et que les animations perdent de leur vivacité. En contraste, les sujets qui ont préférés les composantes activées ont commentés que les animations comportaient moins de sauts de positions et semblaient plus fluides.

Malgré que nous indiquons que ces fonctions sont retenues pour la solution finale, dans une utilisation pratique, l'activation de ces fonctionnalités seraient à la discrétion de l'utilisateur du système ou de l'artiste qui fera les retouches sur l'animation produite.

Les données additionnelles qui ont étés expliqués dans l'article ont également étés ajoutées aux animation de la solution finale. Nous avons montré qu'elles ajoutent beaucoup à la qualité visuelle des animation produite sans influencer la précision de la capture de mouvements. L'ajout de l'information sur les mains et la tête montre un vrai avantage de l'intégration de la cinématique inverse, qui permet d'obtenir des positions et rotations des articulations, même avec de l'information incomplète ou manquante. De plus, le "look at" implémenté pour guider les mouvements de la tête amène des mouvements d'épaules, qui sont complètement statiques dans les animations de références, donnant un air beaucoup plus naturel à l'animation finale.

En résumé, nous retenons que deux caméras donnent un résultat de capture optimal, que l'intégration de la cinématique inverse utilisant la pose courante montre un résultat plus précis, tandis que l'utilisation de la pose précédente ne montre une augmentation de la précision par rapport à aucune cinématique inverse, mais montre aussi une très grande amélioration visuelle à l'animation. C'est donc cette dernière qui sera intégrée dans le test final, afin de tester le pire des cas concernant la précision. De plus, activer les limites humaines a amené une amélioration autant au niveau de la précision et au niveau qualitatif. Cette composante est donc retenue. La limitation de la longueur des os montre de résultats similaires. L'ajout de l'information concernant la tête et les mains sont également retenus, car elle n'impacte pas la précision, mais augmente considérablement la capture. Finalement, même si ce filtrage montre une légère détérioration au niveau de la précision de la capture, celle-ci augmente la qualité visuelle et diminue le nombre de pics dans les graphiques de façon significative et sera également intégrée au test final.

Nous observons des résultats significatifs. En effet, toutes les mesures montrent une amélioration par rapport à la capture faite avec seulement une caméra de profondeur et l'analyse faite par la bibliothèque de détection de squelette. La limite supérieure d'erreur de position est réduite à 9.8 cm, soit une amélioration de 30.9%. La limite inférieure d'erreur est réduite à 3.1 cm, soit une amélioration de 47.7%. Finalement, le nombre de pics est réduit de 7 à 4 et le nombre de déviations par rapport à la capture effectuée par le système FOB est réduit de 3 à 2.

En plus, les tests à l'aveuglette ont été menés avec le même groupe de 66 individus. Ils ont tous choisi la capture de mouvement effectuée avec la technique suggérée et aucun n'a montré de préférence, à aucun niveau, pour celle effectuée par simple caméra 3D.

Ces résultats sont significatifs et montrent que l'intégration de plusieurs caméras de profondeur et l'intégration de la cinématique inverse dans un contexte de capture de mouvement permettent d'améliorer de façon importante les séquences de capture de mouvements produites.

6.2.2 Limitations

Malgré les améliorations amenées à la capture de mouvement par caméra 3D, certaines limitations freinent toujours son utilisation à des fins pratiques.

La première vient de la caméra 3D elle-même. Par sa faible résolution en profondeur et son très haut niveau d'incertitude sur les positions des joints du squelette, une capture de mouvement qui rivalise la précision d'un système commercial avec marqueur est impossible. De plus, le bruit engendré par l'interférence entre plusieurs de ces caméras limite leur utilité en grand nombre. En effet, dans le meilleur des cas, nous observons une erreur minimale de 4.3cm sur la position d'un joint donné. Dans le pire des cas, cette incertitude peut monter jusqu'à 13cm ! En comparant ces valeurs au système de capture de mouvement utilisé comme référence, qui a une erreur maximale de 2.5mm, les limites du système deviennent évidentes.

De plus, les bibliothèques d'analyse de squelette des caméras 3D ne fournissent pas beaucoup d'information par rapport aux joints du sujet capturé. En effet, un squelette limité a 16 joints est disponible des bibliothèques Kinect SDK et OpenNI. De plus l'information par rapport aux rotations des joints disponibles est presque toujours invalide, donc inutilisable pour la capture de mouvement. Par contre, l'avantage de la solution proposée est que l'utilisation de la bibliothèque de cinématique inverse HumanIK d'Autodesk permet de compléter cette information afin d'être utilisée dans un cadre pratique.

La quantité limitée de joints disponible requiert donc une étape de plus dans la capture afin de déduire l'information manquante. Tel que décrit dans les sections précédentes, les informations de positionnement de la tête et des doigts du sujet sont déduites de différentes façons. Ces informations sont limitées et ne peuvent pas représenter les positions réelles du sujet durant la capture. Elles sont donc présentes que pour améliorer la qualité visuelle de l'animation générée.

Enfin, le raffinement des techniques de capture d'images de profondeur permettra l'extraction de l'information squelettique de plus en plus précise. L'amélioration de l'information squelettique diminuera l'effet de correction qu'aura la bibliothèque HumanIK et son influence sera ainsi moins ressentie. En effet, dans l'approche suggérée, le plus grand avantage pratique de l'intégration de la bibliothèque de cinématique inverse est que celle-ci fournit les rotations des os même si l'information est manquante du squelette extrait des images de profondeur. Tant et aussi longtemps que cette information est manquante, l'utilisation de HumanIK demeure pertinente.

6.2.3 Analyse pratique

La méthode de capture de mouvement présentée cherche à améliorer les résultats obtenus à partir de capture basée sur les caméras 3D dans le but d'être utilisée à des fins pratiques. Les caméras 3D ont un coût relativement bas comparé à des systèmes de capture de mouvements couramment utilisés. Ils sont aussi moins intrusifs.

Leur utilité montre donc plusieurs avantages. Par contre, ils ne sont pas utilisés dans l'industrie parce que leur précision laisse à désirer, l'information sur la capture fournie est insuffisante et leur instabilité génère beaucoup d'imperfections dans les animations générées. Leur qualité doit donc être grandement améliorée.

Les résultats discutés dans la section précédente montrent que les animations créées avec la technique de capture suggérée améliorent, autant quantitativement que qualitativement, les résultats obtenus avec une simple caméra 3D. Il est donc nécessaire d'évaluer leur qualité et leur utilisation dans un pipeline de production pratique.

Ainsi, les animations résultantes de la capture de mouvement ont été exportées en format FBX, afin d'être utilisées dans des logiciels de production populaires, telles que Maya et *Motion Builder*. Maya est le logiciel de modélisation 3D le plus populaire, offrant une très grande gamme de fonctionnalités incluant plusieurs centrées sur l'animation. *Motion Builder* est le logiciel le plus utilisé comme solutions de capture de mouvement, et offre plusieurs fonctions utiles, comme la modification et filtrage de courbes, les exportations, etc.

L'importation des animations créées dans *Motion Builder* permet une meilleure visualisation et manipulation. Par exemple, *Motion Builder* permet de filtrer les courbes générées pour diminuer les pics qui se retrouvent dans l'animation originale. L'outil permet également d'interpoler les clés d'animations selon différentes courbes.

Une séquence d'animation générée par la capture de mouvement suggérée a donc été importée dans *Motion Builder* et donnée à un expert de capture de mouvement chez Autodesk Inc. Après avoir filtré les courbes et appliqué une interpolation cubique entre les clés de l'animation, on aperçoit une amélioration significative. Toutefois, selon cet expert, l'animation produite par la technique de capture de mouvement suggérée ne produit pas une animation de qualité suffisante pour être utilisée dans un cadre de production. Toute capture de mouvement demande un certain degré de retouche avant d'être utilisée en production. Dans le cas présenté, le temps et l'effort de l'étape de retouche ne justifient pas l'utilisation du système de capture de mouvement ; il sera presque toujours avantageux de créer l'animation sans capture de mouvement.

Le système suggéré n'est donc pas suffisamment mature pour être utilisé dans le cadre d'une production commerciale. Par contre, les techniques peuvent être réutilisées avec les prochaines générations de caméras 3D qui sauront certainement améliorer leur précision et amener plus d'information aux squelettes.

Par contre, comme suggérée précédemment, l'intégration de la bibliothèque HumanIK avec les caméras de profondeur amène tout de même un grand avantage. La méthode suggérée fournit les rotations associées à chaque joint du modèle humain. C'est cet avantage qui suggère tout de même un potentiel d'utilisation de la méthode de façon pratique. En effet, dans un environnement de production, les animations squelettiques sont représentées avec des rotations par rapport au squelette dans sa pose de base. Cette pose est souvent la pose «T-Stance» pour un être humain, qui montre le sujet avec les pieds à la largeur des épaules, les bras étendus horizontalement de part et d'autre de son corps, avec les pouces qui pointent vers l'avant. Donc avec seulement l'information incomplète des rotations venant des squelettes des caméras 3D, les rotations doivent être déduites, une approche peu pratique et sujette à de grandes erreurs. Avec la méthode produite, les animations peuvent être directement produites et utilisées dans une application multimédia, augmentant grandement la possibilité d'adoption des caméras 3D en tant que solution de capture de mouvements.

6.3 Technique de capture

Malgré le manque de précision et la quantité de bruit trop élevée dans les animations créées à partir de la méthode de capture de mouvement par caméra 3D suggérée, celles-ci montrent plusieurs avantages par rapport aux autres techniques. Avec l'amélioration certaine des technologies de caméras 3D et de leur logiciel d'extraction de squelettes, la précision et qualité de la technique suggérée sera également améliorée.

6.3.1 Avantages et inconvénients

Les inconvénients de l'utilisation de la technique suggérée sont évidents. La précision de l'animation résultante n'est pas suffisante pour justifier son utilisation dans un milieu pratique, même si la technique améliore significativement les résultats d'une simple caméra 3D. De plus, les imperfections dans les courbes d'animations, telles que les pics qui résultent en de très fortes variations de mouvements, mettent en évidence certains mouvements qui ne sont pas plaisants à l'œil.

Par contre, les méthodes de capture de mouvements sans marqueurs montrent certains avantages significatifs par rapport à leur contrepartie. Tout d'abord, les sujets ne sont pas encombrés par des marqueurs ou senseurs qui peuvent gêner leurs mouvements ou même empêcher certaines positions.

Aussi, la technique suggérée est beaucoup moins dispendieuse comparée aux systèmes commerciaux de capture de mouvements. Le matériel nécessaire pour la capture par caméra 3D suggérée est limité à trois caméras Kinect (environ 150$ chacune) et un poste de travail. L'installation du système est également très rapide et demande peu d'espace. Les systèmes de capture de mouvements commerciaux avec senseurs demandent généralement un investissement important pour le matériel (généralement au-delà de 10 000$), une installation permanente et un grand espace pour la capture.

La technique suggérée montre également des avantages par rapport à l'utilisation d'une unique caméra 3D. Elle améliore significativement la précision et la qualité de l'animation produite.

Ces avantages font que l'utilisation de ce système de capture de mouvement serait très avantageuse pour les studios de jeux indépendants, qui ont souvent un budget limité.

6.3.2 Utilisations pratiques

En supposant une amélioration significative de la précision des caméras 3D dans le futur, la technique de capture de mouvement présentée a le potentiel d'être utilisée dans plusieurs industries.

Tout d'abord, l'industrie du jeu vidéo et du film pourrait remplacer leurs systèmes de capture de mouvements pour intégrer des systèmes comme celui suggéré dans ce travail. Ceci mènerait à des cycles de développement plus rapides et à des coûts de production diminués.

De plus, une technique de capture de mouvement rapide, non intrusive et sans marqueur peut grandement bénéficier les applications multimédias interactives autres que les jeux vidéo. On pourrait penser à un écran interactif dans un kiosque de ventes qui détecte un utilisateur et superpose des produits sur celui-ci. Une technique de capture de mouvement rapide est primordiale et un certain degré de qualité et précision doit être maintenu afin de garder l'interactivité.

CHAPITRE 7 AVANCEMENTS ET TRAVAUX FUTURS

7.1 Introduction

La technologie de caméras 3D a beaucoup augmenté en popularité durant les dernières années et ses applications sont adoptée dans plusieurs domaines. Au fur et à mesure que la précision des capteurs est améliorée, l'extraction du squelette humain se fera de façon plus efficace et les résultats seront d'autant plus intéressants.

7.2 Limites

Les limites courantes de la capture de mouvements par caméra 3D dérivent principalement du manque de précision des caméras 3D. Nous avons démontré que l'intégration d'un modèle de cinématique inverse et de contraintes humaines peut en effet améliorer significativement les séquences capturées, mais que celles-ci demeurent de qualité insuffisante pour être utilisée dans un pipeline de production efficace.

Certainement, toute capture de mouvement demande une retouche par un artiste. L'intérêt d'une capture des mouvements d'un être humain plutôt que la création d'une animation dans un logiciel de création de contenus est que la majorité du travail est déjà fait. Les limites d'utilisation pratique sont ainsi définies par le temps qu'un artiste doit consacrer après la capture pour obtenir un résultat satisfaisant.

Aussi, les limites sur le nombre de caméras 3D utilisé ont été étudiés. La limite trouvée, soit trois caméras simultanées, n'est pas conforme avec les résultats d'autres recherches. Ainsi, une étude de matériaux utilisés, de la nature de l'arrière plan et de l'effet de l'éclairage devrait être faite afin de reproduire la limite de 5 caméras trouvée par la recherche mentionnée. Aussi, il serait intéressant d'étudier l'effet d'un matériel rétroréfléchissant dans la qualitée du signal capté par les caméras Kinects.

En améliorant la précision des caméras 3D, du modèle squelettique et du modèle humain y étant associé, ces limites sauront être réduites jusqu'à ce que la solution proposée devienne viable dans un contexte de production.

7.3 Pertinence des recherches concurrentes

Les recherches concurrentes visant à améliorer la capture de mouvement par caméra de profondeur visent à améliorer la qualité du squelette obtenu à partir des images de profondeur. Des modèles de reconnaissances de poses permettent de reconnaître plus fidèlement une quantité variée de poses et d'en déduire des poses intermédiaires en temps réel.

De plus, l'avènement de capteurs actifs avec une meilleure résolution synergise très bien avec les méthodes de reconnaissance de poses en leur amenant plus d'information sur lesquelles basées leurs décisions.

Certains capteurs de profondeur passifs sont également une source d'amélioration potentielle. En effet, avec un capteur passif, la limite de caméras de profondeur n'est plus pertinente et une très grande quantité de capteurs peuvent être utilisés simultanément et leurs informations combinées.

La recherche concurrente la plus pertinente à la méthode proposée est sans doute celle décrite dans Shum and Ho (2012). Celle-ci montre également des résultats de capture de mouvement significativement améliorés. En effet la technique a comme conséquence d'éliminer presque totalement le bruit causé par l'incertitude des mesures des caméras 3D. De plus, un modèle physique est habituellement la méthode utilisée dans les jeux vidéo pour intégrer la cinématique inverse aux modèles 3D. La combinaison de la méthode décrite dans l'article pourrait très bien être intégrée avec la recherche proposée et les plus grandes limites des deux recherches seraient ainsi vaincues.

7.4 Intégration et améliorations

Un des avantages importants de la technique proposée vient de l'indépendance de la méthode de capture et d'extraction du squelette par rapport à l'intégration de la cinématique inverse dans la solution. Ainsi, l'intégration de recherche concurrente et future par rapport à la détection du squelette humain avec une caméra de profondeur peut se faire sans heurt.

Au fur et à mesure que l'efficacité de l'extraction du squelette humain à partir d'images de profondeur augmentera, les résultats qui dérivent de cette recherche s'améliorera en conséquence. Toutefois, une amélioration linéaire n'est pas attendue puisque, comme mentionné précédemment, des améliorations à la technique de capture peuvent chevaucher les corrections faites avec la cinématique inverse. Dans ce cas, la section couverte de la recherche présentée deviendrait obsolète.

La technique proposée peut être améliorée de multiples façons. En effet, les techniques d'intégration de la cinématique inverse explorées dans cette recherche visaient toutes à intégrer le modèle humain après la détection des squelettes et leur fusion en un squelette central. La possibilité d'appliquer un modèle de cinématique inverse directement au niveau de la détection pourrait être explorée, avant même la fusion des données.

Enfin, la fusion des squelettes elle-même n'était pas le point de convergence de la recherche. Par exemple, une approche plus intelligente de sélection de l'espace de transformation des squelettes pourrait mener à de meilleurs squelettes. Une extraction du squelette avec les images combinées, plutôt qu'utiliser chaque image et squelette séparé, serait aussi une amélioration importante qui aurait le potentiel de contrer le bruit émis par l'utilisation de multiples caméras 3D.

CHAPITRE 8 CONCLUSION

La recherche présentée soulevait qu'il est possible d'améliorer les résultats de la capture de mouvements résultant d'une caméra de profondeur en combinant plusieurs de celles-ci avec un système de cinématique inverse adapté au corps humain.

8.1 Synthèse des travaux

Afin d'arriver aux bonnes conclusions vis-à-vis l'hypothèse émise, une méthodologie rigoureuse a été décrite. Celle-ci comprend le développement de plusieurs modules indépendants. Le premier d'entre eux permet l'analyse des effets de la combinaison des données squelettiques en un modèle centralisé.

Ensuite, la cinématique inverse a été introduite dans le système à l'aide de la bibliothèque HumanIK d'Autodesk. Celle-ci a été intégrée à plusieurs niveaux afin de mesurer ses effets et optimiser les conséquences positives sur l'animation finale produite. Nous avons découvert que d'utiliser la pose courante comme entrée de données produit de meilleurs résultats que d'utiliser la pose calculée dans la trame précédente.

Les différentes fonctions offertes par HumanIK provenant de la cinématique inverse appliquée au modèle du corps humain et implémentées par la bibliothèque ont ensuite été testées. Les résultats quantitatifs montrent que les limites appliquées et corrections apportées au squelette résultant de la capture par caméra 3D augmentent la précision de la capture de mouvements. Par contre, les résultats qualitatifs soulèvent une opinion mixte, très dépendante du contexte et de la nature de la capture.

Finalement, de l'information supplémentaire par rapport aux mains et à la tête du sujet a été extraite des images de profondeur. Ces informations ont été injectées dans la bibliothèque HumanIK afin d'augmenter considérablement la qualité visuelle et le réalisme des séquences produites.

8.2 Limitations de la solution proposée

La méthode proposée augmente considérablement la qualité visuelle des animations produites et mène à une précision accrue dans les mouvements capturés. Malgré cette amélioration, la capture de mouvements par caméra de profondeur doit devenir plus mature avant de trouver une place définitive au sein de l'industrie.

Les recherches futures, l’amélioration de technologies de caméras de profondeur et l’intégration de nouvelles méthodes et technologies à leur utilisation quotidienne montrent certainement un futur prometteur face à la technologie et son application pour la capture de mouvements.

8.3 Améliorations futures

La pertinence d’intégrer un modèle du corps humain a été démontrée. Bien sûr, plusieurs améliorations peuvent être amenées à la méthode proposée. En effets, des recherches peuvent être conduites afin de déterminer l’efficacité d’intégrer la cinématique inverse au cœur de l’extraction du squelette par la caméra 3D elle-même. La fréquence et l’intensité de pics de mouvements dans les graphiques produits doivent également être réduites afin de produire des animations utiles dans un contexte pratique.

Finalement, les résultats positifs obtenus et la possibilité d’améliorations à la technique proposée par cette recherche sont encourageants et montrent le potentiel de l’utilisation de caméras 3D dans le contexte de capture de mouvements.

RÉFÉRENCES

Autodesk, “Human inverse kinematics”, 2012, [Online ; accessed 07-July-2012]. En ligne : `http://gameware.autodesk.com/humanik`

S. W. Bailey et B. Bodenheimer, “A Comparison of motion Capture data Recorded From a Vicon System and a Microsoft Kinect Sensor”, Vanderbilt University, Rapp. tech., 2012.

K. Berger, K. Ruhl, Y. Schroeder, C. Bruemmer, A. Scholz, et M. Magnor, “Markerless Motion Capture using multiple Color-Depth Sensors”, dans *Vision, Modeling, and Visualization (VMV), Berlin, Germany, Oct. 4–6*, Oct. 2011, pp. 317–324.

P. Documentation, “Point cloud library (pcl) 1.7.0 - transformationestimationsvd”, 2012, "[Online ; accessed 5-November-2012]. En ligne : `http://docs.pointclouds.org/trunk/classpcl_1_1registration_1_1_transformation_estimation_s_v_d.html"`

B. Hasler, N.and Rosenhahn, T. Thormahlen, M. Wand, J. Gall, et H.-P. Seidel, “Markerless Motion Capture with unsynchronized moving cameras”, dans *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 20-25 June 2009*, June 2009, pp. 224–231.

W. Hinchman, “Kinect for Windows SDK beta vs OpenNI”, 2011, [Online ; accessed 05-November-2012]. En ligne : `http://labs.vectorform.com/2011/06/windows-kinect-sdk-vs-openni-2/`

C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages”, *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, January 2004. DOI : `10.1016/j.ijforecast.2003.09.015`. En ligne : `http://www.sciencedirect.com/science/article/B6V92-4BJVV07-3/2/75e73118cfeba36d23df13dd9c445f3e`

D. F. Huber et M. Hebert, “Fully automatic registration of multiple 3d data sets”, *Image and Vision Computing*, vol. 21, no. 7, pp. 637–650, July 2003.

K. Khoshelham, “Accuracy analysis of kinect depth data”, dans *International Society for Photogrammetry and Remote Sensing (ISPRS)), Zurich, Switzerland, Aug. 25-Sep. 1, 2011.*, vol. Volume XXXVIII-5/W12, 2011, ISPRS Calgary 2011 Workshop, 29-31 August 2011, Calgary, Canada.

A. Lorusso, D. W. Eggert, et R. B. Fisher, “A Comparison of Four Algorithms for Estimating 3D Rigid Transformations”, *MACHINE VISION AND APPLICATIONS*, vol. 9, no. 5-6, pp. 272–290, 1997.

A. Maimone et H. Fuchs, “Reducing interference between multiple structured light depth sensors using motion”, dans *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, March 2012, pp. 51–54. DOI : `10.1109/VR.2012.6180879`

Microsoft Corp, “Kinect for Xbox360”, 2012.

PCL, “Point cloud library”, 2012, [Online; accessed 21-July-2012]. En ligne : `http://pointclouds.org/`

PrimeSense, “OpenNI”, 2012.

Y. Schröder, A. Scholz, K. Berger, K. Ruhl, S. Guthe, et M. Magnor, “Multiple kinect studies”, Computer Graphics Lab, TU Braunschweig, Rapp. tech. 09-15, Oct. 2011.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, et A. Blake, “Real-Time Human Pose Recognition in Parts from Single Depth Images”, dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1297–1304.

H. Shum et E. S. Ho, “Real-time physical modelling of character movements with microsoft kinect”, dans *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, série VRST ’12. New York, NY, USA : ACM, 2012, pp. 17–24. DOI : `10.1145/2407336.2407340`. En ligne : `http://doi.acm.org/10.1145/2407336.2407340`

E. A. Suma, B. Lange, A. S. Rizzo, D. M. Krum, et M. Bolas, “FAAST : The Flexible Action and Articulated Skeleton Toolkit”, dans *2011 IEEE VIRTUAL REALITY CONFERENCE (VR)*, série Proceedings of the IEEE Virtual Reality Annual International Symposium, Hirose, M and Lok, B and Majumder, A and Schmalstieg, D, éd. IEEE; IEEE Visualizat & Graph Tech Comm (VGTC); IEEE Comp Soc, 2011, pp. 247–248, IEEE Virtual Reality Conference (VR), Singapore, SINGAPORE, MAR 19-23, 2011.

A. Sundaresan et R. Chellappa, “Markerless motion capture using multiple cameras”, dans *Computer Vision for Interactive and Intelligent Environment, 2005*, nov. 2005, pp. 15 – 26. DOI : `10.1109/CVIIE.2005.13`

P. R. Winters, “Forecasting sales by exponentially weighted moving averages”, *Management Science*, vol. 6, no. 3, pp. 324–342, 1960. DOI : `10.1287/mnsc.6.3.324`

Z. Zhang, “Microsoft kinect sensor and its effect”, *MultiMedia, IEEE*, vol. 19, no. 2, pp. 4–10, Feb 2012. DOI : `10.1109/MMUL.2012.24`