UQAC
Université du Québec
à Chicoutimi

L'œil synthétique de la machine : la perception temporelle des systèmes de capture de mouvement (MOCAP)

par : David Hurtubise

Mémoire présentée à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de Maître ès arts (M. A.) en art (concentration design numérique)

Québec, Canada

RÉSUMÉ

En proposant une solution utilisant certains procédés issus des technologies de l'intelligence artificielle *Open Source* pour la capture de mouvement, ce mémoire examine certains paramètres qui définissent les performances de la vision artificielle synthétique, notamment les mécanismes d'apprentissage et la latence. Dans le contexte actuel de la prolifération de nouveaux métavers, reproduire la capture de mouvement de l'œil et la synthétiser avec un ordinateur nécessite des connaissances dans les sous-domaines de la vision par ordinateur tels que l'apprentissage automatique, l'intelligence artificielle, le traitement du transfert de données et l'animation 3D. C'est en étudiant les paramètres de la latence que nous avons hypothétiquement établi que ce champ de recherche converge vers une seule et grande question : Dans les systèmes de capture de mouvement industriels, comment fonctionne, sur le plan temporel, l'œil de la machine? Basée principalement sur l'analyse des technologies de la production virtuelle les plus utilisées, la solution *Open Source* proposée dans le cadre de ce mémoire permet de pointer et de mieux évaluer certaines limites du système perceptif informatique par rapport aux solutions existantes. En définitive, ce sont les dimensions qualitatives de la vision humaine qui devraient prédominer sur les systèmes de mesure de la vision et du mouvement par ordinateur. C'est le point de départ des démarches industrielles dont les objectifs reposent d'abord et avant tout sur la production d'effets narratifs.

Mots-clés : Vision par ordinateur, Intelligence artificielle, apprentissage automatique, Capture de mouvement (MOCAP), Reciblage en temps réel, Latence.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

COMPUTER VISION : De l'anglais, La vision par ordinateur est un domaine scientifique et une branche de l'intelligence artificielle qui traite la façon dont les ordinateurs peuvent acquérir une compréhension de haut niveau à partir d'images ou de vidéos numériques. Du point de vue de l'ingénierie, il cherche à comprendre et à automatiser les tâches que le système visuel humain peut effectuer.

XR: De l'anglais, Extended Reality

VP: De l'anglais, Virtual Production

VR: De l'anglais, Virtual Reality

IN CAMERA VFX: De l'anglais, résultat d'une nouvelle méthodologie pour la prise de vue d'effets visuels en temps réel lors d'un tournage de film d'action en direct. Cette technique repose sur un mélange d'éclairage LED, de suivi de caméra en direct et de rendu en temps réel avec projection hors axe pour créer une intégration transparente entre les acteurs de premier plan et les arrière-plans virtuels. Son objectif principal est de supprimer le besoin de composition sur écran vert pour produire des résultats de pixels finaux à l'appareil photo. L'un des défis de la production d'effets visuels en temps réel de haute qualité est de synchroniser la technologie pour tout exécuter simultanément.

MOCAP: De l'anglais, Motion Capture

DOF: De l'anglais, Degree Of Freedom

TF: De l'anglais, Tensor Flow. Artificial Intelligence Library.

AI: De l'anglais, Artificial Intelligence

ML: De l'anglais, Machine Learning

ROM: De l'anglais, Range of Motion

LBE: De l'anglais, Location Based Entertainment. C'est un terme généralement utilisé pour décrire toute forme de divertissement qui se déroule dans un lieu spécifique en dehors du domicile de l'utilisateur - souvent dans un centre de divertissement familial.

METAVERSE: Le terme est régulièrement utilisé pour décrire une future version d'Internet où des espaces virtuels, persistants et partagés sont accessibles via un univers en 3D et 4D. Une définition alternative affirme qu'il s'agit de l'ensemble des mondes virtuels connectés à Internet, qui sont perçus en réalité augmentée ou directement sous la description du WEB 3. Ce concept a été décrit la première fois dans le roman Le Samouraï virtuel, paru en 1992, de Neal Stephenson.

DÉDICACE

Je dédie ce mémoire à ma fille, Gloria Hurtubise, alias mon rayon de soleil.

REMERCIEMENTS

J'aimerais préciser que j'ai réussi à accomplir ce mémoire avec une méthode de travail qui pourrait s'inspirer du chaos. C'est pour cette raison que je tente d'extirper le maximum des données de recherche de mon expérience de vie professionnelle. En écrivant ce mémoire, j'écris également une infime partie de mon histoire. Et j'aimerais garder en souvenir tous ces gens qui ont été de passage dans ma vie et qui m'ont aidé à traverser la tempête.

À vrai dire, des étapes personnelles bouleversantes de ma vie m'ont guidé vers des idées professionnelles, des projets et des recherches technologiques qui se positionnent entre l'intuition et la logique. Terminer ce mémoire ; c'est ma façon à moi de me libérer d'une certaine étape de ma vie. Quoi que j'en vois aujourd'hui un parcours ou un destin plus précis, il m'est arrivé à de nombreux moments de mon existence de ne plus rien y comprendre... Des flashs de création et des moments illogiques ont résulté en des concepts professionnels définis. De ce chaos, est née cette expérience de recherche création. Et qui, j'espère, pourra aider de nombreux chercheurs créatifs ou designers futurs à poser un regard lucide sur leur vision de la technologie.

Je remercie les gens de l'École NAD, soit mon directeur de recherche Yan Breuleux, pour sa passion et son dévouement en tant qu'enseignant mais également en tant qu'ami sincère. Et Christian Beauchesne, pour la chance qu'il m'a offert de croire en la maîtrise.

Je remercie les gens d'Epic Games, Juan Gomez, Kim Libreri, Matt Madden, Sébastien Miglio, Marc Petit, pour la chance que j'ai eu d'apprendre avec l'une des meilleures entreprises en développement de jeux vidéo et de logiciels. Ils m'ont permis de débattre des enjeux théoriques et pratiques du jeu vidéo avec, selon-moi, des personnes extraordinaires. Je remercie les gens de MELS Studios, tels que Martin Carrier, Richard Cormier et Nicolas Fournier, pour l'opportunité de développement et d'apprentissage en studio de cinéma avec la production virtuelle.

Je remercie mes frères, Mathieu Hurtubise et Vincent Hurtubise, pour m'avoir soutenu tout au long de ce projet. Malgré les hauts et les bas ; la fraternité avant tout. Je remercie et embrasse ma fille, Gloria Hurtubise, qui fut ma lumière qui m'a toujours guidé à travers la noirceur. Et je remercie sa maman, Marie-Christine Tétreault.

Je remercie et respecte mon père, Alain Hurtubise. Et je remercie tout spécialement ma sœur, Izabel Hurtubise et ma mère, Jocelyne Martin, mes anges gardiens, qui m'ont guidé dans ce long et difficile pèlerinage.

AVANT-PROPOS

La technologie de capture de mouvement m'a toujours fasciné, du plus loin que je me rappelle. Malgré mon parcours, c'est quelque chose qui est restée et que j'ai toujours continué à analyser, qu'il s'agisse de la technologie de suivi de mouvement (Sensors Technologies), de l'intelligence artificielle (AI), de Tensor Flow (Machine learning), de la latence entre toutes les étapes de rendu en temps réel (Signal Processing) ou des rendus de ciblages d'avatars 3D (Computer Graphics).

Avec ces idées en tête, j'étudiais les systèmes de capture de mouvement (MOCAP), de latence et de ciblage 3D des animations en temps réel avec des engins de jeux tel que Unreal Engine 4 afin de recréer un prototype de capture de mouvement pour la production virtuelle. En 2015, notre compagnie familiale *HB Picture* prit de l'expansion et devint mère de la compagnie Drive VFX basée à Magog, qui a obtenu un programme d'aide en recherche avec le PARI-CNRC, en collaboration avec des étudiants de l'Université de Sherbrooke. Il s'agissait de recherche et de développement sur le positionnement des données des caméras Prime 13 de Optitrack vers le moteur de jeux, le début de Unreal Engine 4 et de la connexion des données de captures en temps réel avec le casque en réalité virtuelle, Oculus DK1. Le sujet du projet PARI-CNRC était de recréer des avatars en temps réel pour une expérience LBE.

Cette recherche nous a permis de créer un plugin UE4 de capture VR, bien avant les accès SDK de Oculus, Optitrack et Unreal. Un projet et des pistes de recherches que nous avons étudiés de 2015 à 2016. Nous étions que junior dans le domaine de recherche, sans vraiment connaître les directions précises pour la programmation. Je me rappelle à cette époque avoir même de la difficulté à différencier une librairie *Dynamic-link Library* (DLL) et un *Software Development Kit* (SDK). Mais bizarrement, malgré les manques flagrants de connaissances en programmation, l'idée était ancrée dans ma tête à la suite de l'analyse des résultats de cette recherche. Par contre, il est faux d'écrire que cette aventure de compagnies en 2016 s'est terminée positivement. Quelques mois plus tard, ces compagnies ont toutes fait faillite. Ce qui m'a entraîné dans un chaos personnel des plus incohérents et illogique… Le pire moment de ma vie si je puis dire… Et j'ai perdu plusieurs choses que j'aimais dans la vie : Ma famille, mon couple, mes amis... C'est d'ailleurs ce qui a poussé mon choix à quitter les cantons de l'est pour m'installer à Montréal.

Après des mois de difficultés financières, que je limite ici de décrire, je fus intégré par chance dans le système académique pour une maîtrise en art à l'école des arts numériques, de l'animation et du design (École NAD) à Montréal, suivant les conseils de Christian Beauchesne et de Yan Breuleux. Ce qui fut le début de la formation de mes idées d'innovations. Je commençai les études, toujours poussé par le désir de continuer les recherches faites avec le PARI-CNRC en 2015 sur la capture de mouvement utilisée avec Oculus Rift. Ce qui m'a poussé à réaliser une première phase de recherche dans le cadre d'un financement Mitacs avec l'UQAC intitulé l'Optimisation du suivi des acteurs dans un espace partagé en réalité virtuelle pour des installations en contexte muséa*l*. Le projet consistait en la création d'un prototype fonctionnel conçu pour une installation immersive pour le musée de l'ingéniosité J. Armand Bombardier. Nous parlons ici d'études faites jusqu'en 2017.

Ce type de projet permettait à l'organisme d'évaluer la pertinence de l'usage d'expériences immersives partagées utilisant des capteurs numériques (Sensor Technology) en VR et explorant la position des mains dans l'environnement pour la prise d'objets virtuels. De plus, cette recherche avait pour objectif de contribuer aux problématiques liées à la simulation de présence et de téléprésence de plusieurs acteurs au sein d'un environnement de réalité virtuelle (Computer Graphics) partagé. L'objectif était d'augmenter le niveau de captation physique de plusieurs acteurs et par le fait même, la physicalité de ceux-ci dans le monde virtuel. Ce projet tentait de résoudre un certain nombre de défis techniques de locomotion associés à l'usage des technologies des casques VR à cette époque.

J’ai par la suite continué le travail chez Neweb Lab, en 2017, comme Directeur Technique. Ce qui m’a permis de comprendre également le “pipeline” des artistes 3D chez la compagnie mère, Vox Populi, sur l’émission : “*Et Dieu créa Laflaque*”. Il s’agissait à cette époque, pour la première fois, de créer un schéma MOCAP, selon le pipeline d’effets visuels 3D post procédural. Un premier schéma qui regroupait les domaines des capteurs MOCAP et du *Computer Graphic* avant l’intégration AI.

C’est à la fin de l’année 2017 que j’ai intégré la compagnie Epic Games (Unreal Engine 4) en tant que gestionnaire de compte technique, avec un premier *White Paper* officiel en tant qu’auteur sur un pipeline de VFX 3D, précisant plusieurs sous domaines de l’infographie 3D, soit : *ICI Laflaque Broadcast quality within a tight timeline*. J’ai travaillé 3 ans avec Epic Games et Unreal Engine sur de nombreux projets de production virtuelle, ayant aidé des compagnies comme Netflix, Digital Dimension, Animal Logic, DreamWorks, MELS Studios, Moment Factory à développer des pipelines de studio avec Unreal Engine 4 et de nouvelles techniques de production virtuelles.

C’est au cours de mon expérience chez Epic Games que j’ai également eu la chance de collaborer sur le développement des nouveaux outils de productions virtuelles *In Camera VFX*. J’ai également eu la chance de présenter deux *Tech Talk* à Siggraph avec Epic Games, dont un sur la réalité mixte en 2018 et l’autre sur la Production Virtuelle en 2019 avec les artistes techniques dévoués du jeu vidéo *Fortnite*. Je suis finalement promu en tant que *Solution Architect* en 2020, sous l’équipe de Juan Gomez. Humblement, je peux dire que j’ai appris beaucoup de Epic Games et des talents professionnels que j’ai pu côtoyer là-bas. En fin de l’année 2020, j'ai quitté Epic Games par choix pour devenir Directeur de la Technologie de production Virtuelle chez MELS Studios. Nous avons travaillé sur de nombreux films en production virtuelle tels que *Transformer 7* et *Disappointment Boulevard*. Continuant les recherches en étroite collaboration avec Epic Games sur la production virtuelle, nous avons continué et continuons à ce jour avec MELS de faire de la recherche et développement pour la production virtuelle avec volume LED. Nous avons d’ailleurs reçu un *MEGA GRANT* de Epic Games en 2021. Cette étape de ma vie m’a permis d'apprendre davantage sur les différentes saveurs de la production virtuelle. De fait, comprendre les diagrammes technologiques de la production virtuelle *In Camera VFX* pour d’importantes productions cinématographiques et d’en apprendre plus sur ce type d’installation qui retrouve sa complexité dans les nombreuses connexions logicielles, matérielles et intergicielles. D’ailleurs, c’est précisément à cette époque que la similarité entre les schémas des volumes LED, des casques VR, des volumes de captures (MOCAP) et de l'œil humain, m’est venue à l’esprit et de les classer en paramètres symbiotiques. Entre autres, relater la familiarité entre la latence et la précision de n’importe quelle de ces architectures biologiques vs virtuo-logiques.

INTRODUCTION

Dès 2012, j'ai pris connaissance des méthodes d'estimation de capture de mouvement issues de la biomécanique, plus spécialement pour l'animation 3D. À cette époque j'avais deux passions : le cinéma et les technologies en biomécanique. Et un équilibre entre ces thèmes commençait à se matérialiser par l'étude des pionniers de l'époque comme Zemeckis ou Cameron. Principalement, mon intérêt se dirigeait sur des travaux de recherche réalisés sur l'une des productions virtuelles contemporaines les plus marquantes : *Avatar* (Cameron, 2009). Dans ce film, le réalisateur, James Cameron, s'est appuyé sur les approches utilisées dans la recherche sur la capture de mouvement du film *The Polar Express* (Zemeckis, 2004) pour développer un nouveau pipeline de production cinématographique qui propose une relation en temps réel entre l'environnement physique et virtuel. Cette nouvelle approche de production permet au réalisateur de voir la performance d'animation d'un acteur dans un environnement virtuel en temps réel. (Ng, 2012; Thacker, 2012a).

La production virtuelle est un concept relativement nouveau qui couvre une large sélection d'approches axée sur la combinaison de contenu physique et virtuel en temps réel. Dans le domaine du divertissement, les productions virtuelles commencent à dominer l'industrie en ce qui concerne la création d'images de synthèse. Les effets visuels et la production virtuelle sont des domaines dans l'industrie cinématographique qui ont tendance à se chevaucher.

Actuellement, sur le marché des technologies du domaine de la production virtuelle, il existe de nombreuses solutions de réalité virtuelle capables de capturer et de suivre tous les mouvements d'un corps humain. Et la plupart des marques de réalité virtuelle bien connues proposent leur propre système. Que ce soit *Optitrack, Vicon, Xsens, Rokoko* ou des ensembles VR comme *Oculus* ou *HTC Vive*, ces marques proposent un nombre croissant de capteurs de mouvements pour MOCAP et d'intégrations de logiciels de solution de suivi. En revanche, la plupart de ces systèmes de réalité virtuelle intégrés n'offrent pas beaucoup d'outils pour étendre le calibrage et le reciblage

(de l'anglais *retargeting),* entre leur système de capture et l'animation d'avatars virtuels 3D en temps réel. Chacun de ces systèmes comprend une approximation des mouvements de l'utilisateur grâce à l'utilisation d'une série d'étalonnages prédéfinis et d'amplitude de mouvement du *range of motion* (ROM). Et l'utilisation de logiciels tiers comme *MotionBuilder* d'Autodesk pour améliorer la qualité de reciblage de l'animation 3D est toujours très nécessaire dans l'industrie professionnelle des effets visuels et des films.

Ce mémoire est basé sur l'examen de nombreuses solutions industrielles pour l'industrie de la production virtuelle. Les données sont basées sur de multiples expériences professionnelles intégrant, en tant qu'artiste technique, des connaissances provenant des sous-domaines de la vision par ordinateur telles que les technologies d'analyse de capture de mouvement, l'intelligence artificielle, l'apprentissage automatique, l'animation 3D et le traitement du signal.

Dans l'ensemble des solutions de captation de mouvement, la position de l'humain et de l'avatar 3D ne peut pas "naturellement" se produire dans un espace virtuel car les lois normales de la physique réelle ne s'appliquent pas comme dans le monde virtuel. Cela peut être l'aspect le plus difficile à évaluer. C'est-à-dire comprendre les liens relatifs entre la réalité et le virtuel. C'est d'ailleurs le problème central de la vision par ordinateur. L'ordinateur doit traduire les lois de notre monde physique et dimensionnel et les cibler dans un espace virtuel. Sans les multiples correspondances entre le monde physique et virtuel, il n'y a pas d'interaction standard ou virtuelle avec l'hôte humain et son avatar 3D. Pour pouvoir préciser la position complexe des polygones 3D de l'avatar, des mesures doivent être faites entre le monde physique et virtuel, l'hôte et l'avatar 3D. Les coordonnées des polygones d'un avatar doivent être spécifiées par rapport à l'origine de l'hôte lui-même. Le lien entre l'objet physique et virtuel signifie que le moteur 3D doit recalculer les coordonnées des points des polygones par rapport à l'origine physique réelle de l'hôte.

Créée par Rudolf Laban, *Laban Movement Analysis* (LMA) (1879-1958) est une méthode créative d'étude du mouvement pour observer, décrire, noter et interpréter le mouvement humain.

Un ancien chorégraphe de danse nommé Rudolf Laban a utilisé un système pour enregistrer les mouvements des danseurs. La notation de Laban décrit une séquence fixe de pas, des danses prescrites, fixée par des règles.

Le système LMA donne un aperçu du style de mouvement personnel et augmente la prise de conscience des séquences corporelles. Principalement, Laban définit l'analyse de mouvement en 4 catégories :

- L'effort
- Le corps
- La forme
- L'espace

Le système de Laban est la base de tous les étalonnages ROM qui font partie de tous les principaux systèmes de capture de mouvement MOCAP connus aujourd'hui. Pour la plupart des systèmes MOCAP conventionnels, il est impératif de suivre une suite logique de calibrage de pose pour permettre au logiciel et au système de capture de valider l'effort, le corps, la forme et l'espace. Plus précisément et selon les différents systèmes, l'acteur hôte peut également avoir besoin de suivre une série de mouvements ROM ou de positions de ses membres afin de spécifier l'axe des articulations telles que les coudes, les épaules, les hanches et les genoux. Ces axes communs sont à la base de la définition de la plupart des avatars 3D dans les logiciels de MOCAP conventionnels.

Mais, malgré l'utilisation de ROM pour estimer l'avatar 3D, les systèmes MOCAP présentent des problèmes généraux tels que la latence et la précision. Et ces problèmes ont conduit à de nombreuses recherches et développements dans les domaines de la vision par ordinateur, de la production virtuelle et des effets visuels.

Pour répondre à ces problématiques, il faut d'abord, dans le premier chapitre, définir comment les systèmes actuels perçoivent les acteurs et objets. Il s'agit de comprendre quels sont les paramètres qui définissent les mesures à l'œil humain pour une architecture de production virtuelle. Dans le premier chapitre de ce mémoire, j'exposerai une mise en contexte du MOCAP, du *Computer Vision* et de la relativité avec la vision humaine, je présenterai ensuite ses paramètres, leurs relations et leurs facteurs avec les technologies MOCAP dans l'industrie des effets spéciaux et de la production virtuelle. Entre la présentation des domaines de la vision de synthèse et de l'analyse encore plus synthétique des différents systèmes de captures, cela nous mènera à l'exposé de la question et l'objectif de recherche.

Afin de répondre à la question posée et atteindre l'objectif, le deuxième chapitre explique la méthodologie utilisée dans cette recherche. Selon une méthode itérative, nous évaluons divers prototypes selon les paramètres définis :

- De la latence
- Des paramètres d'apprentissage
- Du reciblage 3D en temps réel

Le troisième chapitre exposera l'analyse des différents prototypes qui m'ont permis de préciser mon exploration des moyens de transfert des biomécaniques humaines telles que la vision humaine en relation à la vision symbiotique ; en temps réel et avec plusieurs solutions MOCAP jusqu'à faire des prototypes itératifs d'animation de capture de mouvement avec des solutions de simples caméras *red, green and blue* (RGB) ou d'algorithmes de AI et d’apprentissage automatique.

Le quatrième et dernier chapitre, consacré à la discussion et l'analyse des résultats, traitera du potentiel et des limites de ces technologies en relation avec la présente recherche, les problématiques et la méthodologie utilisée.

CHAPITRE 1

LA PROBLÉMATIQUE DE LA SYNTHÉTISATION DES MOUVEMENTS PAR ORDINATEUR

1.1 MISE EN CONTEXTE DE LA RECHERCHE

Dans ce premier chapitre, nous définissons comment les systèmes actuels perçoivent les acteurs et objets. Il s'agit de comprendre quels sont les paramètres qui définissent les mesures de la capture de mouvement pour une architecture de production virtuelle. Dans le premier chapitre de ce mémoire, j'exposerai une mise en contexte du MOCAP, de la vision par ordinateur et de la relativité avec la vision humaine, je présenterai ensuite ses paramètres, leurs relations et leurs facteurs avec les technologies MOCAP dans l'industrie des effets spéciaux et de la production virtuelle. Entre la présentation des domaines de la vision de synthèse et de l'analyse encore plus synthétique des différents systèmes de captures, cela nous mènera à l'exposé de la question et l'objectif de recherche.

1.2 DÉFINITION ET BREF HISTORIQUE DE LA TECHNOLOGIE MOCAP

L'étude de la locomotion et les procédés d'observations de l'animation se sont développés au fil de l'histoire, bien avant la création technologique des systèmes de capture de mouvement MOCAP. Cette étude a pris plusieurs formes et les termes ont évolués, tel que le décrit cette citation:

> *“Le philosophe grec Aristote (-383 à -321) a publié, outre de nombreux autres ouvrages fondamentaux, un texte (bref) ΕΡ Ι ΠΟΡ ΕΙΑS ΖΩΙΩΝ sur la démarche des animaux. Il a défini la locomotion comme "les parties qui sont utiles aux animaux pour le mouvement en place"... une jambe mène au côté droit d'un triangle rectangle. Comme les légendes sont égales alors, celle au repos doit se plier... au genou... Ce texte est le premier document connu sur la biomécanique. Il contient déjà, par exemple, de nombreuses observations sur les schémas de mouvement de l'homme lorsqu'il est impliqué dans une activité particulière.” (Klette, Tee, 2004)*

Historiquement, ces recherches ont intéressé bon nombre d'artistes. Léonard de Vinci (1452-1519) a écrit dans ses carnets de croquis qu'il était indispensable pour un peintre de visualiser totalement l'anatomie des nerfs, des os, des muscles et des tendons, de sorte qu'il

comprenne leurs mouvements et leurs contraintes, par exemple : quel tendon ou quel muscle provoquent un mouvement particulier chez l'homme.

Sans faire l'histoire complète de la question, il est utile de rappeler certains faits marquants. Il y a plus de 100 ans, en 1915, le caricaturiste Max Fleischer a inventé le rotoscope pour faciliter la production de films d'animation. L'appareil projetait des films d'action en direct sur une table lumineuse, une image à la fois. Le designer traçait ensuite l'image projetée sur du papier. Il était ensuite assemblé pour réaliser des films d'animation. Plusieurs décennies plus tard, au début des années 1960, Lee Harrison III a créé le premier corps de capture de mouvement équipé de potentiomètres et de personnages 3D animés sur un moniteur *cathode ray tube* (CRT) en temps réel. Plusieurs courts métrages ont été réalisés au cours de la période en utilisant cette technologie. Finalement, Harrison a remporté un *Emmy Awards* pour ses contributions techniques. Dans les années 1980, Ginsberg et Maxwell ont créé un système de capture d'animation qui utilisait des *light emitting diode* (LED) clignotantes attachées à un acteur entouré d'un ensemble de caméras. Des caméras triangulaient la position 3D des LED en temps réel. La vague actuelle de *Deep Learning* a commencé avec Alex Krizhevesky. Celui-ci a remporté le défi de reconnaissance visuelle à grande échelle ImageNet (ILSVRC) en 2012.

En ce qui concerne le sujet de recherche de ce mémoire en relation avec mon expertise, il est nécessaire de mentionner quelques expériences fondatrices. En 2015, la société Epic Games, créatrice du moteur de jeu Unreal Engine, a rendu possible la création d'une expérience de réalité mixte appelée *The Void*. Afin de contrôler un avatar par le corps de l'hôte dans un monde de réalité mixte, l'expérience est décrite comme suit : l'hôte porte un casque VR, un capteur de suivi de la main, une combinaison de capteurs infrarouges et un ordinateur pour alimenter le casque. Les joueurs peuvent alors participer collectivement à un jeu vidéo multi-utilisateurs synchronisant l'espace réel et virtualisant leur corps en mouvement avec un avatar 3D sans utiliser de contrôleurs et avec moins de capteurs que le MOCAP traditionnel.

En 2018, Epic Games s'est associé à *3Lateral*, *Cubic Motion*, *Tencent* et *Vicon* pour créer l'une des séquences les plus complètes du suivi de l'animation finale jamais réalisées en temps réel. Dans le projet "*Siren*", un personnage en temps réel haute-fidélité était piloté par une combinaison MOCAP et un véritable hôte. C'était le début de l'humain numérique en temps réel.

Ces dernières années, la vulgarisation de l'apprentissage automatique a inspiré de nouvelles approches basées sur l'apprentissage pour résoudre le problème de l'estimation et de l'animation de la pose humaine. On peut parler ici des Méta humains. Et Epic Games a lancé en 2021, le premier regard sur *Meta Human Creator*, une nouvelle application basée sur internet pour créer des humains numériques entièrement animés et combinant certains outils d'IA pour une animation de visage réaliste. Et d'ailleurs en 2022, une version de Keanu Reeves créée avec un Méta humain 3D de l'application de Epic Games et utilisant la technologie de *Deepfake*, apparaît sur une chaîne Youtube de *Not Face Video Studio*. Les auteurs de la chaîne spécialisée avaient réussi à créer une superbe version numérique de l'acteur Keanu Reeves, qui était d'un réalisme impressionnant.

Avec l'avènement de processeurs et de graphismes puissants, on peut désormais parler de concepts de vision par ordinateur qui conservent d'énormes quantités de données, de réseaux de neurones et reproduisent plus précisément la vision synthétique. Et le processus d'apprentissage automatique en vision par ordinateur a gagné en popularité en termes de recherche de solutions à des problèmes apparemment simples mais compliqués tels que la classification, l'évaluation et l'estimation d'images.

Au cours des prochaines sections, il sera question de la manière dont les systèmes de captation perçoivent les objets et les corps.

1.2.1 L'ÉVOLUTION DES SYSTÈMES MOCAP CONVENTIONNELS

La plupart des systèmes MOCAP corporels conventionnels ont évolué en deux types de classes : soit les systèmes optiques ou inertiels. Les capteurs et marqueurs optiques, comme leur nom l'indiquent, utilisent la lumière pour capturer les données de mouvement.

Les marqueurs optiques passifs réfléchissent la lumière, tandis que les capteurs optiques actifs transmettent la lumière infrarouge (IR) pour fournir des signaux aux caméras ou aux stations de base. Lorsque plusieurs acteurs sont impliqués, des marqueurs actifs peuvent être configurés pour transmettre un identifiant différent pour chaque acteur. Les marqueurs actifs aident le serveur à séparer les données de mouvement provenant de différents acteurs.

Pour suivre efficacement le mouvement, les marqueurs passifs et actifs reposent sur la conception et le placement de corps rigides, qui sont des configurations fixes (ou « constellations ») de trois marqueurs ou plus. Le système de suivi utilise les distances connues entre ces marqueurs pour aider à suivre le mouvement. Un corps rigide de marqueurs passifs repose sur les emplacements relatifs des marqueurs dans le motif, tandis que les corps rigides des marqueurs actifs reposent directement sur le motif lumineux.

Les systèmes optiques n'ont en réalité qu'une seule limite sérieuse : la caméra ou la station de base doit être capable de « voir » un capteur afin d'obtenir sa position. Cela signifie que les personnages proches les uns des autres sont susceptibles de bloquer les capteurs les uns des autres, ce qui fait que le système "perd de vue" les membres ou le torse du personnage de temps en temps. Plusieurs caméras ou stations de base peuvent aider, mais ne peuvent pas éliminer complètement le risque d'occlusion.

Inversement, les capteurs inertiels utilisent des gyroscopes, des accéléromètres et des magnétomètres pour calculer la position actuelle du capteur dans l'espace XYZ et transmettre ces

données à une station de base, avec un signal qui contourne les joueurs et les obstacles.

Les systèmes inertiels ont également une limitation sérieuse : étant donné que la position XYZ du joueur est déduite de données telles que l'accélération du capteur plutôt que de l'emplacement physique du capteur lui-même, les données peuvent être légèrement inexactes. Un tel écart peut facilement s'amplifier et faire dériver l'avatar de l'emplacement réel du joueur, entraînant des problèmes d'emplacement des avatars ne correspondant pas à l'emplacement du joueur dans l'espace physique.

En raison de la limitation des systèmes inertiels, la plupart des captures de production virtuelle ont évolué vers des systèmes de capture de mouvement optique et qui utilisent des mécanismes logiciels qui déduisent les positions des joueurs lorsque les marqueurs sont masqués par les caméras.

Comme pour tout système de capture de mouvement, l'étalonnage de l'hôte sera nécessaire avant chaque session MOCAP. Chaque système a sa propre méthode d'étalonnage utilisant une série de ROM. Pour effectuer un bon ciblage des poses, il est nécessaire d’avoir un nombre considérable de capteurs.

1.2.2 LA PRODUCTION VIRTUELLE ET LE MOCAP

À l'époque, et encore aujourd'hui pour de nombreuses raisons, l'architecture du MOCAP pour *visual effects* (VFX) traditionnel tel que décrit dans l’itération 1 est encore utilisée dans la plupart des pipelines de production virtuelle. La production virtuelle, ou In-Camera VFX est une nouvelle méthodologie qui a vu le jour en 2020, avec la venue des technologies d’écrans LED géants. Les sites de tournages deviennent alors des immersions en temps réel, utilisant les écrans LED pour produire des environnements digitaux en arrière-fond, lors du tournage des acteurs. Bien qu’en 1941, Orson Wells utilisait déjà une technique similaire pour capturer des images en utilisant

un dispositif d'image rotatif à manivelle, c'est plutôt avec l'avènement des technologies d'immersions en 3d temps réels que Epic Games et ILM présentèrent les premiers concepts de la production virtuelle, dans la série du Mandalorian. La technique principale de tournage de production virtuelle exige d'utiliser un système d'engin de jeu 3D en temps réel tel que Unreal Engine, pour permettre de capturer un décor 3D sur écran LED qui interagit avec les acteurs en temps réel. C'est précisément lors de la fusion des données du monde réel et des données du monde virtuelles pour terminer en format In-Camera VFX, qu'est utilisé le système MOCAP de capture de mouvement pour permettre de comprendre en temps réel la position de la caméra en lien à l'espace virtuel qui est digitalisé sur le mur. Cette technique repose sur un mélange d'éclairage, de suivi de caméra en direct et de rendu en temps réel avec projection hors axe pour créer une intégration transparente entre les acteurs de premier plan et les arrière-plans virtuels.

1.2.3 FACEBOOK ET LA CAPTATION DE MOUVEMENT IA

L'utilisation de l'intelligence artificielle pour la capture de mouvement MOCAP a fait ses débuts dans la production virtuelle avec des casques VR et des solutions AR. Par exemple, la solution la plus courante consiste à fournir simplement un casque tel que l'Oculus Quest, et des contrôleurs, pour reproduire un avatar 3D de la personne. L'année 2012 a été marquée par le retour en force de la réalité virtuelle. C'est au cours de cette année que la société Oculus a lancé une campagne Kickstarter avec un casque de réalité virtuelle révolutionnaire. En 2014, la société a été rachetée par Facebook pour 2 milliards de dollars. L'apparition de ce nouveau marché a engendré une prolifération de nouveaux casques VR tels que : Sony PlayStation VR, Samsung VR Gear, HTC Vive, etc.

Cette nouvelle ruée vers l'or virtuelle a amené tout le secteur des parcs à thèmes virtuels ou en des termes techniques, le secteur du Local Based Entertainment (LBE). L'expérience LBE est un cousin de l'expérience de réalité virtuelle à domicile utilisant seulement le casque et les contrôleurs mais dans un contexte collectif et utilisant le lieu réel comme support à l'espace virtuel.

En principe, une expérience LBE s'efforce d'obtenir une véritable immersion ; se rapprochant plutôt du concept des métavers. Ces expériences LBE procurent un espace fixe et ouvert avec de la place pour plusieurs joueurs, un équipement (MOCAP) et des indices physiques comme des accessoires.

D'ailleurs, il existe des différences importantes et nettes qui concernent les portées relatives de ces deux types de divertissement : VR à domicile et VR LBE. Et ces données se rapprochent des problématiques des espaces virtuels à multi-acteurs. En 2015, la compagnie The VOID a d'ailleurs tenté de complètement intégrer les idées de l'immersion totale de l'hôte vers l'avatar 3D pour des productions à grands espaces et à plusieurs acteurs.

Le matériel nécessaire à l'expérience LBE, peut se composer de plusieurs serveurs, de cartes graphiques haut de gamme, d'ordinateurs en sac à dos, de capteurs optiques tels que *Optitrack* et d'autres matériels que les consommateurs n'auraient pas normalement à la maison. Tout grand espace pour des expériences de genre LBE nécessite des mécanismes de suivi placés à intervalles autour de la pièce. L'espace VR peut être rendu beaucoup plus grand que l'espace physique grâce à des environnements 3D intelligemment conçus et à l'utilisation de portails virtuels. Lorsque l'on parle en général de VR, les espaces de plus de 50 pieds x 50 pieds deviennent des systèmes de capture plus puissants que les solutions de capture existantes comme Oculus ou HTC Vive. Plus le nombre de joueurs est élevé, plus les exigences en matériel et en serveur sont élevées. Bien que de nombreuses configurations sont possibles, la plupart des recommandations pour une expérience LBE sont d'utiliser une solution MOCAP externe avec des caméras optiques, un ordinateur serveur principal, une machine MOCAP dédiée, des casques HMD et des sacs à dos ordinateurs qui servent de clients. Sans compter tous les différents logiciels et extensions pour combiner tous ces matériels et logiciels. Dans une pièce plus petite, on peut utiliser une solution relativement peu coûteuse telle que le système VIVE Lighthouse, qui peut gérer la capture de mouvement dans un espace allant jusqu'à 900 pieds carrés. La combinaison du casque VIVE avec cinq trackers VIVE donne une configuration décente pour une configuration de suivi complet du

corps 6DoF (six degrés de liberté). Nous avons d'ailleurs vu émerger des systèmes miniatures de LBE tel que Hologate en 2011, qui encore à ce jour, utilisent les systèmes de VIVE.

Outre le système de casque VIVE, il existe également le concurrent direct, Oculus. Celui-ci se base sur des paramètres similaires. En 2017, Facebook a fixé l'objectif en rachetant Oculus d'obtenir non seulement un suivi de IK (Inverse Kinematic) intégré dynamique, mais aussi d'optimiser le processus de cartographie avec son système *simultaneous localization and mapping* (SLAM). Son système devait fonctionner plus facilement, n'importe où, et pas seulement dans une scène avec des limites définies par des balises ou autres outils, le tout avec une latence sous les 10 millisecondes (ms).

La réponse à cette question est arrivée par l'hybridation du système normal de capture de mouvement Oculus avec le système SLAM de cartographie IA[1].

Alors que les concurrents tels que HTC VIVE, utilisent des équipements invasifs tels des caméras et des *trackers*, l'approche de Facebook innove en présentant une solution hybride avec le système SLAM. La localisation et la cartographie simultanées SLAM réside dans le problème informatique consistant à construire ou à mettre à jour une carte d'un environnement inconnu tout en gardant simultanément une trace de l'animation du corps de la personne.

Les changements au niveau du IA se font à grande vitesse et d'autres méthodes telles que le *Neural Radiance Fields* (NeRF) voient le jour et connectent ensemble ce que l'on pourrait appeler l'imagination de l'ordinateur et la vision de l'ordinateur par les systèmes de réseaux antagonistes génératifs (GAN). Cette méthode permet de créer des scènes entières à l'intérieur d'un réseau neuronal, à partir de photos statiques. Ces systèmes IA transcendent les effets d'animation 3D par le processus d'édition d'images basées par l'apprentissage automatique. Ces images 2D sont

[1] Voir: [https://mixed-news.com/en/meta-shows-stunning-full-body-tracking-only-via-quest-headset/], 2 octobre 2022.

combinées aux informations que nous avons en 3D. Dans ce contexte, nous pouvons parler du procédé de *Deepfake* qui est apparu en 2017 et qui a pris le monde au dépourvu sur le forum Reddit, en présentant de fausses vidéos d'acteurs connus, créées avec cette solution d'encodage numérique. Des vidéos qui sont modifiées par l'IA et des codes d'apprentissage automatique et dont le code a été déposé sur Github. Depuis, plusieurs versions de logiciels *Deepfake* ont vu le jour tel que DeepfaceLive et DeepfaceLab. On a pu percevoir en 2020, dans la série de Star Wars, Le Mandalorien, l'utilisation du Deep Fake pour recréer le jeune Luke Skywalker. Ces *Deepfake* sont des médias synthétiques dans lesquels une personne dans une image ou une vidéo existante est remplacée par un acteur tel que Mark Hamill de Star Wars, en tirant parti de puissantes techniques d'apprentissage automatique et de IA.

Dans ce contexte, ce mémoire se situe donc du côté d'une meilleure compréhension des systèmes de capture de mouvement nécessitant un minimum d'appareillage et la combinaison par celles-ci avec des solutions IA.

Dans le cadre de ce mémoire, nous avons concentré notre attention sur deux paramètres principaux: *la latence et l'apprentissage*. Ces deux éléments sont interdépendants dans la mesure où l'apprentissage doit s'effectuer dans une temporalité propre à la perception temporelle de l'utilisateur. C'est le ciblage en temps réel. En d'autres termes, les paramètres d'apprentissage, pour être invisibles, doivent s'effectuer en suivant des durées imperceptibles pour l'expérience utilisateur. C'est ce que nous allons analyser

1.3 LES PARAMÈTRES D'ANALYSE DU MOUVEMENT

Puisque la capture de mouvement est un domaine de connaissance très large, il est nécessaire de définir les paramètres de la capture de mouvement et d'étudier les systèmes de captation menant à la question de recherche.

1.3.1 LA LATENCE

Il est important de faire la distinction entre la latence et la fréquence d'images :

- La latence est une mesure de temps nécessaire pour que les données de mouvement du corps du joueur au système, qui à son tour déterminent la vitesse à laquelle la vue dans chaque casque est mise à jour. La vitesse de transfert des données dépend en grande partie des capacités du système MOCAP choisi, mais peut également être affectée par la vitesse de votre réseau et vos paramètres dans Unreal Engine. La latence est exprimée en millisecondes (ms). Plus elle est faible, mieux c'est.

- La fréquence d'images est la vitesse à laquelle chaque image des visuels est rafraîchie (affichée) sur un écran. La fréquence d'images dépend en grande partie des capacités du *hardware*, mais peut être affectée par la complexité de la scène (vitesse de rendu dans Unreal Engine). La fréquence d'images est transmise en images par seconde (FPS). Plus elle est élevée, mieux c'est.

Examinons un exemple simple du mouvement d'un hôte capturé lors d'une séance (MOCAP) de production virtuelle. Les *trackers* sur l'hôte transmettent un flux de données sur la position du corps dans l'espace et la direction des membres et des axes à chaque instant consécutif. Si la latence est faible, chaque donnée reflète la position de l'hôte en quelques millisecondes. Puis l'ordinateur restitue l'avatar 3D et reproduit la réalité virtuelle sur l'écran, les écrans LED par exemple. Une latence élevée entraînera une inadéquation entre la réalité et le virtuel. En général, la vision par ordinateur maximale nécessite une latence ne dépassant pas 8 ms puisque l'œil humain traite une image toutes les 13 à 20 millisecondes, selon une parution dans le journal du MIT par Anne Trafton en 2014 (*In the blink of an eye*).

ÉQUATION 1 : Latence de l'œil humain à la vision par ordinateur

Œil humain : traite une image toutes les 13 à 20 millisecondes

Computer Vision : nécessite une latence inférieure à 10ms.

Exemple : pour les écrans, le débit recommandé pour ce que l'on présente à l'utilisateur sur un écran est de 90 fps.

SI fps = 90 images / 1 seconde

ALORS ms = 90 trames / 1000 ms

1000 ms / 90 trames = 11,11 ms / trame

SORTIE = 11,11 ms / trame

Cet équation nous indique qu'une fréquence d'images de 90 fps se traduit par la production d'une nouvelle image toutes les 11,11 millisecondes, ce qui est inférieur à la plage de 13 à 20 ms pour l'œil humain. Et qui est inférieur à 10 ms pour la vision par ordinateur. Pour atteindre une telle fréquence d'images, un système de capture de mouvement en temps réel utilise deux ensembles de données :

- Les données en direct sont transmises au réseau à partir de capteurs (MOCAP) sur le corps hôte.
- Les données prédictives sont calculées par le système. Il "remplit les blancs" pour répliquer les données en direct d'un utilisateur afin de montrer l'animation de son avatar à un autre utilisateur.

Les données en direct doivent transiter vers le serveur où elles peuvent ensuite être analysées, avec une latence mesurée en millisecondes. Les données prédictives, en revanche, sont calculées dans un logiciel, par exemple, Unreal Engine - la latence des données prédictives est généralement d'une fraction de milliseconde. L'utilisation d'une combinaison de données en direct et prédictives

adoucit le mouvement de l'avatar 3D et offre généralement une meilleure expérience visuelle à l'utilisateur. Ce mélange de données est appelé dans ce contexte : la réplication. Il donne le spectre complet de la latence du matériel au logiciel.

1.3.2 LES PARAMÈTRES D'APPRENTISSAGE

La position de l'objet et la reproduction 3D ne peuvent pas "naturellement" se produire dans un espace virtuel car les lois normales de la physique réelles ne s'appliquent pas comme dans le monde virtuel. Il est nécessaire de faire des ajustements à la position du corps physique la plupart du temps. Cela peut être l'aspect le plus difficile de la performance relative entre la réalité et le virtuel. Sans les paramètres d'apprentissage de la pose, les soit-disant références entre le monde physique et virtuel, il n'y a pas de standard dans la production virtuelle. Ici, le message principal est défini dans la mesure.

Pour pouvoir spécifier la position complexe des polygones 3d de l'hôte, des mesures doivent être faites entre le monde physique et virtuel. Les coordonnées des polygones d'un objet doivent être spécifiées par rapport à l'origine de l'objet lui-même, soit un sommet, soit le centre de gravité. Le lien entre l'objet physique et virtuel signifie que le moteur 3D doit recalculer les coordonnées des sommets des polygones en respectant l'origine physique réelle et l'origine du point zéro (Origine) de chaque monde.

Avec les solutions de MOCAP traditionnelles, ces paramètres d'apprentissage sont réalisés à l'aide de la série de ROM définies par LABAN. Bien que vous ayez probablement vu des images d'acteurs de cinéma portant des combinaisons de MOCAP ajustées comme on peut le voir dans les captures de mouvement du film Avatar et Avatar 2 de James Cameron, de telles combinaisons sont quelque chose que de nombreux chercheurs tentent de supprimer dans les travaux futurs sur la capture de mouvement en production virtuelle. Nous pouvons repenser aux travaux de Facebook sur SLAM. L'évolution du MOCAP avec l'intelligence artificielle fait partie de l'évolution en

automatisant la précision des ROM sans avoir besoin de combinaisons et de nombreux *trackers* sur l'hôte. Et sans avoir besoin d'un gros processus d'étalonnage. Dans les paramètres d'apprentissage de l'IA, nous utilisons un modèle d'apprentissage automatique qui est une représentation mathématique d'un processus du monde réel. On peut représenter un modèle d'apprentissage par 3 couches qui définissent finalement ce modèle :

- La couche 1 est un ajout de données natives
- La couche 2 est le calcul des mises à niveau mathématiques en fonction des informations archivées des données natives
- La couche 3 est le résultat

On peut classer les différentes solutions Open Source IA disponibles par ce tableau :

Tableau 1 : Regroupement des techniques d'IA pour la vision par ordinateur.

Computer Vision Libraries	*Machine Learning Libraries*	*Deep Learning Models libraries*	*Deep Learning Framework*
- Open CV - DLib - OpenFace - OpenBR - Pytorch - Keras - YOLO	- Tensor Flow - MediaPipe - OpenPose - PoseNet - Coco - Human 3.6	- Deep Mocap DMC 2.5D - DeepfaceLab - DeepfaceLive	- SSD - Caffe - DensePose

1.3.3 LE PROBLÈME DU RECIBLAGE EN TEMPS RÉEL

Le reciblage est le processus qui consiste à prendre l'animation d'un personnage et à l'appliquer à un autre. Si vous affichez des avatars 3D, vous devrez cibler le mouvement capturé de chaque hôte suivi vers son avatar.

La plupart des problèmes proviennent de la mise en correspondance d'un vecteur 3D correspondant parfaitement au mouvement ou à la pose biomécanique. L'estimation de la pose en 3D utilise généralement un modèle de corps humain comme représentation paramétrable d'un corps réel. La complexité de ces modèles de corps peut fortement varier, selon le scénario d'application voulu. Dans de nombreux cas, un modèle de squelette virtuel est utilisé, composé de segments de membres reliés par des articulations. Chaque articulation peut avoir un ou plusieurs angles de flexion et une pose humaine est, dans ce cas, entièrement donnée par la configuration angulaire de toutes les articulations. Les modèles de corps squelettiques peuvent suffire dans plusieurs cas en estimant la pose avec la cinématique inverse (IK), bien connue dans le domaine de l'animation 3D. Malheureusement, une méthode IK ne peut pas toujours résoudre la pose biomécanique, car les poses du corps ne dépendent pas des caractéristiques présentées.

Le reciblage devient délicat lorsque l'hôte et l'avatar ne correspondent pas étroitement en taille et en poids. Une façon de résoudre ce problème est de décomposer l’avatar 3D en plusieurs parties (ex : tête, haut de jambes, bas de jambes, bras, avant bras, poitrine, abdomen), chacune avec une taille et une proportion différentes, et de les faire correspondre avec la position des acteurs dans leur ROM. Une autre approche consiste à cibler les mouvements de l'hôte vers un avatar à taille unique en utilisant un facteur de combinaison de filtres et de mélanges automatiques, basé sur la taille et la proportion de l'hôte. Bien que ce dernier puisse être plus simple à configurer, il peut entraîner des mouvements d'avatar inhabituels.

Si les avatars ne suivent pas la structure du squelette humain (par exemple, une créature des bois avec un os supplémentaire dans sa jambe), le processus de reciblage est encore plus complexe. Quelle que soit l'approche que vous utilisez, vous devrez tester votre solution de reciblage sur des personnes de différentes formes et tailles avant de la considérer prête pour MOCAP. Et la plupart des solutions de reciblage utilisent une combinaison de logiciels tiers qui ajoutent une latence supplémentaire dans le processus.

1.4 QUESTION DE RECHERCHE

Par conséquent, en relation avec le problème du reciblage, notre question de recherche est la suivante:

Pour le reciblage en temps réel, comment rapprocher le plus possible les yeux synthétiques de l'ordinateur des mécanismes de l'œil humain ?

Cette question de recherche se situe entre les technologies de l'*Embodiment* (Stern, 2013) et du *Spatial Computing* (Shekhar, Feiner et Als., 2016), les travaux sur lesquels j'ai travaillé constituant des modèles / prototypes au niveau des relations entre le corps et la captation. Ce qui pourrait se traduire par le terme de la physicalité ou en anglais, *Embodiment.*

La physicalité est un concept utilisé pour décrire l'ensemble des expressions visibles et tangibles d'un acteur. Ou selon Leach (2013), comme la présence matérielle d'un corps dans l'espace. La physicalité est généralement attribuée à la façon dont un acteur se comporte en tant que personnage, ce qui comprend les mouvements, les gestes, les poses et les expressions faciales.

Par conséquent, la physicalité fait référence à l'incarnation du personnage et à la façon dont un acteur utilise son corps pour offrir une performance, tel que l'avait défini LABAN avec les performances de danse. Dans ce cas-ci, l'incarnation du corps biomécanique et ces variables peuvent se référer aux poses ROM d'un acteur vers un avatar dans un environnement virtuel.

Imaginons maintenant les problèmes possibles associés à la physicalité, en précision à la synchronisation biomécanique entre l'acteur du monde physique et l'avatar du monde virtuel. Ces avatars de haute qualité sont une partie importante des reproductions virtuelles fascinantes, que ce

soit pour les films, les jeux ou les effets visuels. Cependant, animer un avatar 3D pour qu'il corresponde au mouvement de son utilisateur est une tâche essentielle difficile.

L'interconnexion et le synchronisme entre engins de rendu, logiciel de capture et systèmes de capteurs de mouvement ajoutent des ralentissements et des contraintes avec tous les processus de calcul, de calibrage et de reciblage de la pose sur l'avatar 3D. En plus des défaillances de synchronisation et des problématiques d'occlusion.

1.5 OBJECTIFS DE LA RECHERCHE

Le premier objectif de cette recherche est de transférer une synthèse de certaines connaissances utiles sur les procédés de captation de mouvements acquise par de nombreuses années de pratique professionnelle vers le milieu industriel et scientifique.

Le deuxième objectif de cette recherche est de proposer, à partir des données du premier objectif, des prototypes de systèmes exploitant certaines possibilités d'apprentissage offertes par les solutions *Open Source* utilisant l'intelligence artificielle.

Globalement, ces deux objectifs visent à opérer un transfert d'expérience du monde industriel vers le secteur académique. Les prototypes du second objectif ont en fait été réalisés dans l'optique de mieux comprendre comment l'ordinateur, du point de vue de la captation du mouvement, perçoit le corps humain.

1.6 HYPOTHÈSE DE RECHERCHE

Bien que vous ayez probablement vu des images d'acteurs de cinéma portant des combinaisons moulantes ajustées avec de nombreuses boules réfléchissantes argentées, de telles combinaisons ne sont pas utiles pour toutes les types de création d’effets visuels et de production

virtuelle. Il existe autant de types de systèmes MOCAP que de logiciels de pipeline pour reproduire une animation 3D avec reproduction des poses d'un acteurs, que ce soit pour une solution traditionnelle ou en temps réel. Sans regrouper ces solutions et les tester, il est difficile d'établir des liens entre leurs différents paramètres et également de savoir comment l'on pourrait se rapprocher d'une synthétisation de l'œil et de la vision du corps par ordinateur.

La prémisse de cette recherche est que la création de différentes itérations et essais de systèmes de MOCAP et pipeline d'animation 3D pour la production virtuelle, permettra de définir les paramètres nécessaires à répondre à notre question de recherche. Et afin de synthétiser la capture de mouvement par ordinateur et de l'automatiser, nous explorerons l'utilisation du IA pour la capture de mouvement en créant une solution *Open Source* de capture de mouvement MOCAP, qui combinera les paramètres et qui sera disponible pour les chercheurs futurs.

CHAPITRE 2

MÉTHODOLOGIE

2.1 RECHERCHE-CRÉATION

Le lien entre création et recherche n’est pas évident. Cette conception de la recherche-création peut difficilement rencontrer la manière dont la recherche se conçoit dans le secteur de la vision par ordinateur. Ce domaine de connaissances provenant de la science informatique repose sur une méthode scientifique rigoureuse où les données sont soumises aux critères de scientificité classique. Le principal reproche envers la recherche-création peut se résumer par la critique de Stévance et Lacasse (2014) qui affirment certains constats tels que: la création n'est pas la recherche, la diffusion artistique n'est pas de la recherche et enfin, que la recherche n'est pas une contemplation de soi. Malgré l'impression d'une certaine critique radicale des auteurs, il faut mentionner qu'il s'agit principalement d'une critique positive. Par exemple, Sophie Lacasse est titulaire d'une chaire de recherche en recherche-création musicale. Les constats visent principalement à délimiter le territoire de la recherche-création.

En ce sens, pour éviter tout malentendu sur les liens unissant art et recherche et sur la logique de la recherche-création, il est nécessaire de rappeler qu’il existe de multiples approches pour développer de nouvelles connaissances. Quivy et Van CAmpenhoudt (1995) n’hésitent pas à nous rappeler que pour faire une véritable recherche systématique, il faut d’abord rompre avec les préjugés et les idées conçues. De plus, en suivant l'examen systématique des méthodes de recherche effectuées par Louis-Claude Paquin et Cynthia Noury, démontrent un véritable foisonnement d'approches de recherche-création, souvent opposées et hétéroclites. Une constante se dégage toutefois, la majorité des modèles : ils explorent les liens entre théorie et pratique.

Notre positionnement est le donc suivant: dans le contexte de cette recherche, en suivant le cadre épistémique de la pratique réflexive de Donald Schön (1983), l'expérience professionnelle

du praticien peut constituer une source de connaissances transférable vers le milieu de la recherche. Ce qui veut dire:

> *"Éclairer les paroles et les gestes des praticiens en explorant ce qui ressort des modèles d'activités spontanées que leur pratique engendre." (Schön, 1996, 36)*

Nous pouvons symboliquement affirmer que le trait d'union entre recherche et création souligne l'activité du praticien, qui est tout à la fois, le nœud et l'ancrage entre ces deux termes. Selon Chapman et Sawchuk, la création elle-même sert de moyen de générer des nouvelles connaissances qui peuvent agir comme une forme de recherche en soi. Globalement, le positionnement de ce mémoire est de concevoir le lien entre recherche et création sous la forme d'un cycle continu et itératif d'apprentissage par la production de prototypes.

2.2 LE CERCLE EXPÉRIENTIEL DE KOLB

Le modèle méthodologique du cercle expérientiel de Kolb (David A. Kolb 1984) repose sur l'expérience en tant que facteur d'apprentissage et de développement. Notre but est de discerner des méthodes efficaces pour les recherches qui se produisent dans une époque de changements radicaux au niveau des technologies numériques. Les professionnels de l'industrie des effets visuels dans le domaine de la captation de mouvement sont confrontés aux changements fréquents et constants des environnements de production. Mais il demeure que les pipelines de production en effets visuels restent lent et traditionnel comparé à celui de la production virtuelle. Un amalgame entre ces 2 mondes doit être observé et redirigé vers une nouvelle structure de recherche-création. Et ainsi donner un nouveau pipeline aux artistes créatifs de la production en temps réel.

Selon le cycle de Kolb, les prototypes et essais nous conduisent à de nouvelles expériences et permettent de suivre les changements technologiques radicaux : le cycle de Kolb peut toujours recommencer. Le cycle de Kolb est une méthode d'apprentissage itérative marquée par les étapes de l'expérience de la réflexivité, de la modélisation et de l'expérimentation. Une

interaction entre un apprenant et un objet qui conduit à une représentation mentale constituant un outil pour comprendre le monde, s'y adapter ou le modifier. En somme, c'est un processus par lequel la connaissance est créée par la transformation de l'expérience.

David A. Kolb (avec Roger Fry) a créé son célèbre modèle sur 4 zones principales d'apprentissage qui permet de classifier les apprenant en divers types sociaux :

1. Expérience concrète - (une nouvelle expérience de la situation est rencontrée, ou une réinterprétation de l'expérience existante).
2. L'observation réflexive (Les incohérences entre l'expérience et la compréhension).
3. Conceptualisation abstraite (La réflexion donne naissance à une idée nouvelle, ou à une modification d'un concept abstrait existant).
4. Expérimentation active. Le *hand's on* (l'apprenant les applique au monde qui l'entoure pour voir les résultats).

Plus précisément, Kolb et Fry (1975) font valoir que le cycle d'apprentissage peut commencer à tout l'un des quatre points - et qu'il devrait vraiment être abordé comme une spirale continue. Kolb (1984) a identifié quatre profils d'apprentissage : Accommodateur, Assimilateur, Convergent et Divergent :

1. Accommodateur : Approche pratique. Sont classés en tant que manipulateurs. Plus basé sur l'intuition. Faire et ressentir.
2. Assimilateurs : Personnes logiques. Conceptualisent par les idées et concepts abstraits et par la pré-observation. Ils assimilent les observations disparates. Ce sont des observateurs et penseurs qui sont basés sur l'approche scientifique.
3. Convergent : Personnes théoriques. Saisissent l'expérience par la compréhension abstraite et la transformation. Par l'action. Penser et agir.

4. Divergents : Ils observent et génèrent des idées. Apprennent mieux à travers l'expérience concrète et l'observation réflexive. Récepteurs et observateurs.

L'inventaire des profils d'apprentissage de Kolb, permet de savoir si un apprenant met l'accent principalement sur l'expérience concrète, l'observation réfléchie, la conceptualisation abstraite ou l'expérimentation active. En ce qui me concerne, je me situe entre la pensée convergente et divergente selon le contexte. C'est pour cette raison que j'ai opté pour une méthodologie reposant sur les récits de pratique.

2.3 ATTEINTE DES OBJECTIFS DE RECHERCHE

Ce mémoire repose sur un récit de pratique de projet (Paquin, 2019) en contexte professionnel visant à dégager de nouveaux savoir sur la capture de mouvement en tant qu'objet de recherche. Les données sont constituées en deux catégories qui forment les objectifs :

Pour l'atteinte du premier objectif, il s'agit de transférer une synthèse de certaines connaissances utiles sur la captation des mouvements acquis par de nombreuses années de pratique en contexte professionnel. Cet objectif se situera dans la mise en forme du pré-terrain : il sera constitué du récit de pratique d'application des technologies de capture de mouvement pour des projets en contexte professionnel. Il permet de positionner la recherche du second objectif.

Le deuxième objectif repose sur le récit de pratique du développement de deux systèmes de capture de mouvement MOCAP industriels utilisant l'intelligence artificielle. Il constitue le terrain de la présente recherche : c'est-à-dire l'analyse du développement de deux itérations d'un système open-source visant à mieux saisir les relations entre les paramètres de latence et d'apprentissage.

2.4 MÉTHODE DE DESIGN ITÉRATIVE

Dans le cadre d'une présentation au Forum MUTEK à Montréal en 2022, j'ai eu la chance de présenter un diagramme de ma démarche de recherche-création. Ce modèle propose une interprétation du cycle de Kolb en intégrant des étapes de recherche et développement. À ce sujet, Louis-Claude Paquin et de Cynthia Noury ont cité:

> *"Henk Borgdorff énonce que le processus créatif est l'instrument de la recherche-création et que le médium de la création est en lui-même le moyen le plus efficace pour articuler, documenter, communiquer et diffuser les résultats de cette recherche. Il précise que si des expressions discursives accompagnent la recherche, celles-ci ne peuvent jamais prendre la place de la « pensée » artistique. Elles peuvent tout au mieux « l'imiter », c'est-à-dire être utilisées dans une construction réflexive post-hoc du processus de recherche-création.Pour évaluer la recherche-création, Tomas Hellström distingue deux valeurs artistiques de l'œuvre pouvant être appréciées l'une par le public et l'autre par les acteurs du cadre institutionnel associé. Il ajoute une troisième forme, une valeur intermédiaire, qui s'incarne dans le commentaire intellectuel produit par les praticiens. Ainsi, la valeur de la composante création de la recherche-création est selon les normes de la critique artistique." (Paquin, Noury, 2018)*

C'est donc pour suivre cette idée de la valeur intermédiaire qu'il m'est venu l'idée de cartographier ma démarche en tant qu' artiste technique ou chercheur-créatif . Ce diagramme d'un pipeline de la recherche-création suit une démarche de design itérative, une méthodologie qui procède par divers prototype successifs (Steinkuehler, et als. 2012). Elle permet d'aborder dans un premier temps, chaque paramètre particulier (temporalité et apprentissage) à partir d'un prototype fonctionnel. Les prototypes sont évalués par des échantillonnages de mesure permettant de rédiger une table des paramètres en analyse des résultats. Une seconde approche qualitative repose sur la synthétisation des informations relatives aux nombreux systèmes de MOCAP. Ceci afin de les catégoriser. La méthode repose sur un cycle qui a caractérisé ma recherche depuis plusieurs années. La phase de l'éveil technologique repose sur l'idée que dans une perspective de recherche-création, il y a une attention particulière non pas aux technologies en soi mais aux nouvelles possibilités qu'elles suggèrent. Ce qui amène à une phase de recherche et développement qui peut aboutir à une création. De nouveaux problèmes surgissent alors en contexte pratique.Ceux-ci révèlent alors la nécessité de chercher de nouvelles connaissances qui

n'avaient pas été imaginées avant la phase de développement. On formulera donc des hypothèses pour aller chercher ces nouvelles connaissances, ce qui nous amène alors à la nécessité de documenter le chemin parcouru et procéder à une synthèse des connaissances. Bien que ne procédant pas toujours par les réseaux scientifiques, les différents récits de pratiques exposés ici ont tous suivi le même cheminement.

CHAPITRE 3

ITÉRATIONS MOCAP EN MILIEU PROFESSIONNEL

On analysera ici quatre itérations se concentrant sur les variables de la temporalité et de l'apprentissage.

3.1 RÉCIT DE PRATIQUE 1 – LE MOCAP TRADITIONNEL POUR VFX

En 2018, j'ai eu la chance de participer et d'écrire sur un changement d'un pipeline traditionnel vers un pipeline en temps réel avec L'équipe de production d'ICI Laflaque, un hebdomadaire animé par des caricatures politiques. Il est important ici de noter que je n'ai été que support technique dans cette production et que de nombreux artistes CG, animateurs et FX étaient nécessaires à cette production.

Mais cela m'a permis de comprendre le flux de travail pour une série de nombreux épisodes de 30 minutes dont chaque épisode devait être produit dans un délai de 7 jours, de l'animation à la sortie finale. Tout cela, en utilisant un pipeline traditionnel avec la post-production. Sans aucune finalité en temps réel.

Dès le départ, le rendu 3D dans un temps rapide était nécessaire pour respecter ce délai de 7 jours, soit une semaine. Pendant 14 ans, Vox Populi Productions a produit ICI Laflaque sur ce calendrier agressif en utilisant le rendu 3D rapide en OpenGL d'Autodesk MotionBuilder pour un rendu final. Après avoir reconnu que Unreal Engine 4 pouvait leur donner une grande amélioration de la qualité visuelle, Vox Populi a décidé de passer à Unreal Engine pour la saison 2018-2019.

Il est intéressant de noter ici que le pipeline de MotionBuilder se rapproche, déjà à cette époque, d'un pipeline en temps réel sans vraiment l'être. Sans avoir le rendu final, la plupart des animations et des captures de mouvement permettent de travailler en temps réel avec MotionBuilder. C'est d'ailleurs avec cette logique que l'équipe de Vox Populi a travaillé avec le

système de capture Phase Space et MotionBuilder de 2004 à 2018 en supervisant la moitié du pipeline de Laflaque dans un semi-temps réel.

Le système Phase Space était un système de capture de mouvement optique actif qui utilisait 34 caméras Fujifilm X-E2. Un système souvent utilisé dans l'industrie MOCAP pour son faible coût versus sa précision. Mais il s'agissait d'un système moins précis que Optitrack ou Vicon. La capture des animations MOCAP de Laflaque contenait une petite quantité de bruit qui devait être corrigée par des artistes MOCAP dans MotionBuilder.

Cette itération de système MOCAP traditionnel a permis de détailler les étapes de création d'avatar 3D et d'animation :

- La position des caméras optiques doit être faite de manière précise. De façon à diriger les volumes de captures des caméras en différents axes de découpage selon la grandeur du volume de capture. Et selon les intersections entre les caméras.
- Chacune des caméras doit être ajustée selon ses paramètres de lentilles tels que le focus et le zoom pour calibrer le volume de capture.
- Une étape dit comme le *wanding*: avec l'aide d'un bâton de calibration avec capteurs prédéfinis en mm doit être fait suivant des mouvements en 8 ou circulaires, dépendamment de la capture optique Passive ou active.
- Un calcul de la position du point centre zéro est fait à l'aide d'un triangle à 3 capteurs prédéfinis en mm. Le point zéro devient le centre de toute chose dans le monde virtuel et la mesure du monde virtuel.
- Il faut installer des capteurs passifs/ actifs sur le costume de l'acteur, environ 49 selon la solution utilisée. 49 capteurs permettant la plupart du temps de définir un corps complet, sans les mains et le visage.
- La position des capteurs avec des points précis de l'articulation de l'acteur est une nécessité pour la qualité de la capture de l'animation humaine.

- La calibration et le reciblage des points de captures de l’acteur en avatar sont faits dans le logiciel à l’aide d’une pose en T.

À cette époque, et encore aujourd'hui pour de nombreuses raisons, la façon la plus courante de créer des avatars animés en 3D consistait à utiliser des logiciels d'animation haut de gamme basée sur des images clés, tels que Maya, 3ds Max ou Générateur de mouvement. Bien que ces systèmes permettent un contrôle précis du mouvement, de la position et de la synchronisation de chaque élément de la scène, les utilisateurs doivent apprendre une suite complexe d'outils et de commandes pour faire fonctionner leurs interfaces à seuil élevé et à plafond élevé. Bâtir l'expertise nécessaire pour utiliser efficacement ces outils demande du temps et de la patience. Une alternative plus simple consiste, comme dans le pipeline de MOCAP de Laflaque, à filmer un artiste entraînant des marionnettes physiques avec son propre corps. La diffusion kinesthésique en temps réel de la marionnette permet au marionnettiste de se concentrer sur son jeu d'acteur. Puis, les animations de l'avatar de l'acteur sont enregistrées et envoyées à l'animateur 3D de métier, qui travaille et corrige les actions demandées en temps réel. De là vient le terme de semi-temps réel. Un pipeline hybride de capture entre le MOCAP temps réel et la post-production.

Durant cette exploration de l'itération 1, il était question d'utiliser au moins 4 ordinateurs sur un réseau, qu'il soit hors ligne ou en ligne sur un protocole *local area network* (LAN) de *transmission control protocol* (TCP) et *internet protocol* (IP), ou de configurations *user datagram protocol* UDP, l'ajout d'un nouvel ordinateur et d'une connexion réseau supplémentaire est synonyme de configuration / problèmes supplémentaires et de latences supplémentaires.

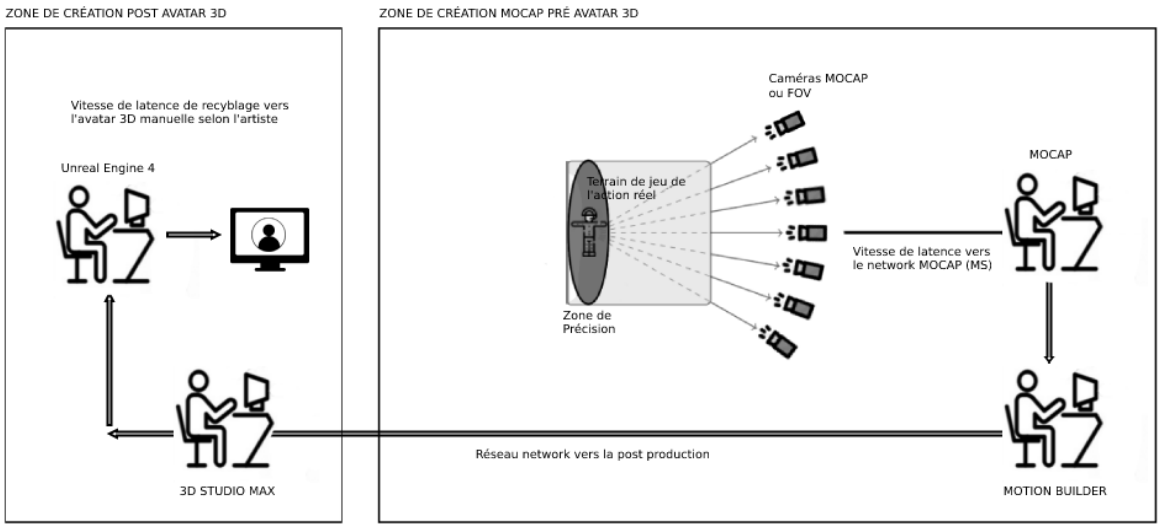

Figure 1 : Diagramme de la création d'un avatar 3D en post production. © David Hurtubise

L'équipe utilisait également le rendu de *MotionBuilder* en OpenGL pour les aperçus et les rendus finaux. Alors que les aperçus pouvaient jouer en temps réel à 30 fps, la sortie finale pouvait être facilement générée à une vitesse inférieure au temps réel de 5 à 10 fps. En utilisant Unreal Engine 4 en 2017, Vox Populi voulait améliorer la qualité et la vitesse de rendu final. La différence entre le 3D rendu de OpenGL et de DirectX n'étant pas négociable.

Dans le passé, entre autres, la recherche avec le PARI-CNRC m'avait fourni un idéal de base pour la conversion de la latence, de la capture au rendu ; soit 8 ms (millisecondes). Une recherche et une analyse des solutions d'interconnectivité entre le rendu en temps réel et la capture de mouvement offrent un effet de présence maximal dans un délai de 8 ms. 8ms équivaut à 120 images par seconde ou 120 fps, ce qui correspond à un multiple de 24 et 30 images par seconde. Motionbuilder limite la fenêtre d'affichage à 60 fps (le plugin live link, limitant le transfert des données du "ViewPort"). En contrepartie, la plupart des matériels optique de MOCAP peuvent atteindre 240 fps. Mais doivent à priori être capturés par leur propre système avant de transférer le data MOCAP vers MOBU.

Cette étape de correction de temps est similaire à l'étape du ciblage décrite dans le chapitre 1. Soit de convertir du data MOCAP à 120 fps de systèmes de captures de Phase Space vers les multiples d'images (FPS) conventionnels de MOBU qui suivent la logique FPS des caméras de production vidéo et qui convertissent vers la norme de FPS pour le rendu TV et cinéma : 24, 29.97, 30, 60 FPS par exemple.

Or, une capture directe avec Optitrack et Motive donnant 240 FPS ne serait pas utilisable en *visual effects* (VFX) sans le reciblage nécessaire vers une logique, image par secondes, relatives à la caméra de production vidéo utilisée dans MOBU.

- 240 / 100 = 24 FPS

Une deuxième problématique est la nécessité de séparer toutes les étapes du pipeline de MOCAP traditionnel avec postproduction en 4 différentes étapes de travail. Impliquant beaucoup de dépendances à des systèmes tiers afin d'obtenir un résultat de rendu d'avatar qui n'était même pas temps réel à cette époque.

1. Logiciel de capture en temps réel
2. Recyclage avec Motionbuilder
3. Correction artistique avec 3D Studio Max
4. Rendu avec Unreal Engine 4

Ce qui devient également un problème important lors des questionnements de ce mémoire afin de créer des avatars 3D en expérience XR. Car chaque logiciel de calculs demande en fait, son lot de calculs et ne fait qu'augmenter le niveau de latence dans le procédé. Déjà lors de l'analyse des résultats de cette première itération, je pouvais affirmer qu'il fallait diminuer le nombre de machines dépendantes au processus de calcul complet.

Pour optimiser leur temps de rendu et leur qualité d'image, l'équipe de Vox Populi gardait un œil sur Unreal Engine depuis quelques années comme solution de rendu en temps réel, impressionnés par les projets R&D tel que Blackbird de The Mill et le projet de Rogue One de ILM, 2 compagnies VFX très connues qui utilisaient Unreal Engine 4 à cette époque.

Pour les personnages de Laflaque, les artistes utilisaient 3ds Max, puis un export vers MotionBuilder et finalement vers Unreal Engine. Nous pouvons définir cette étape par le reciblage définit plus haut dans ce mémoire.

Les corps des personnages étaient "rigged & skinned" avec une combinaison d'os et de modèles ainsi que des objets factices appelés (*null*). Ici, il était important de noter la structure standard de hiérarchie de skeleton entre MotionBuilder et Unreal Engine 4 sans quoi les exports et imports ne marchaient pas.

Des concepts essentiels qui guident et structurent la définition de cette itération. La caractérisation est l'endroit où vous annotez un squelette pour correspondre à un modèle de caractère prédéfini. Dans Motionbuilder, les squelettes avaient et on toujours en 2021 une norme d'utilisation, mais ils sont également flexibles, comparativement à Unreal Engine 4. Ils peuvent être facilement ciblés vers un autre type d'organisation à l'aide d'une étape de caractérisation entre 2 avatars (squelette) virtuels. Cela facilite grandement la configuration de nouveaux personnages. Et cela ajoute un facteur de vitesse d'exécution important lors de la création d'avatars en temps réel. Les contraintes vous permettent d'imposer une certaine logique à la manière dont les articulations des avatars se déplacent suivant les réelles articulations des acteurs dans le vrai monde. Les contraintes doivent être faciles à mettre en place tout en étant flexibles. Ces principes étant d'autant plus importants pour les concepts théoriques du Métavers. Encore une fois, cette étape de reciblage de l'avatar 3d prenait parti des paramètres de variables problématiques édifiées au chapitre 1.

Outre ces facteurs de transferts de poses et d'axes entre Motion Builder et Unreal Engine 4 qui devraient être théoriquement, d'une extrême authenticité pour définir une meilleure corporéité, mais qui ne l'est pas du tout. La nomenclature de base des assets 3D était complètement différente. Menant à de l'incohérence entre les différents logiciels. Chaque personnage avait son propre *namespace* et sa propre hiérarchie de squelette. Par exemple, un personnage dans MOBU nommé "X" pouvait avoir un *namespace* avec deux hiérarchies squelettiques différentes nommées Référence 1 et référence 2.

Cette méthode de délimitation des personnages et des squelettes ne se transportait pas bien dans Unreal Engine 4 pour plusieurs raisons :

- En interne, MotionBuilder stocke ces sous-références avec deux points (:) entre le nom de l'espace de nom et le nom de référence. Cette convention de *namespace* ne fonctionnait pas à l'époque pour l'exportation via FBX, vers Unreal Engine 4.

Les têtes de personnages avaient environ 50 formes de ROMS modélisées avec des Blendshapes. Le format FBX était utilisé pour l'export depuis 3ds Max vers MotionBuilder et ainsi que Unreal Engine 4. À cette époque, plusieurs limitations avec le format FBX diminuaient le matching 1:1 entre les différents exports et imports des assets 3D et des différents logiciels. Tel que le nombre de *bones* du squelette pouvant être utilisé pour les influences des *bones* vers les *skins weights* dans Unreal Engine, qui était limité à 8. Et qui est maintenant défini à 256 par une option dans UE4 et principalement utilisé pour les Meta Human.

3.2 RÉCIT DE PRATIQUE 2 – LE (MOCAP) POUR LA PRODUCTION VIRTUELLE

Le rendu In-Camera VFX est le résultat de la capture par caméra de ce mélange. Son objectif principal est de supprimer le besoin de composition en post-production VFX. Mais, l'un des défis de la production d'effets visuels en temps réel de haute qualité est de synchroniser les

technologies pour tout exécuter simultanément. Et cette étape de la technologie doit se faire en accord avec la latence, les paramètres d'apprentissage et le reciblage de la bonne position de la caméra réelle.

Lors de mon expérience passée chez Epic Games, j'ai beaucoup travaillé avec cette architecture de production virtuelle, avec les solutions Optitrack et Vicon MOCAP, qui sont des trackers optiques, comme le système spatial Phase qui fut utilisé pour Laflaque. C'était la solution utilisée pour créer des cinématiques du jeu vidéo *Fortnite* du concepteur Epic Games, un processus qui a été décrit dans une conférence technique à Siggraph sur la production virtuelle dans UE4 en 2019.

C'est en 2021 que j'ai joint l'équipe de MELS Studios et que j'ai vraiment eu la chance de créer un système de production virtuelle avec LED de toutes pièces. Le système LED de MELS utilisait un mur LED circulaire de 60 pieds de diamètre avec la technologie Roe et Brompton, qui permettait une meilleure vitesse de latence pour la reprojection. Dans cette deuxième itération, il était question de prototyper un système fonctionnel pour la production de tournage cinématographiques tel que Transformer 7. Un système minimisant les contraintes de temps d'opération dites humaines et machines, minimisant les problèmes de synchronisation des GPU internes du module *NDisplay* de Unreal Engine - une technologie qui permet de diffuser plusieurs écrans par le même serveur.

Dans cette itération, il y avait trois problèmes en particulier. Le premier était le développement d'une architecture du système à basse latence et le contrôle avec premier prototype Plugin du streaming de l'image pour faciliter l'opération humaine et logicielle. En ce sens, le système de capture de Optitrack a été choisi pour sa faible latence et son mouvement haute fréquence de 4 ms et 240 fps. Les tests de validations des écrans étant effectués avec une localisation de diverses combinaisons de configurations LED. En tout, une dizaine de tests systèmes ont été fait avec des produits LED comme Novastar et Roe avant de choisir les éléments du prototype de production virtuelle de MELS.

Le deuxième problème était le développement d'une interconnectivité entre le système de capture et le système de production virtuelle avec une latence maximale de 8ms. L'incertitude vient du fait que l'architecture d'un système de production virtuelle contient un nombre élevé de composantes tierces qui sont interconnectées. Ce qui en résulte la plupart du temps à une latence excédant 50 ms, ce qui dépasse beaucoup les contraintes temporelles du projet. Pour se faire, la création d'un logiciel autonome est nécessaire afin de réduire la demande en latence.

Le troisième problème était le rendu et la qualité finale de l'image sur l'écran LED. Il était énormément difficile de reproduire de manière efficace un environnement 3D qui se déplace à la vitesse de l'œil synthétique, soit dans ce système, la caméra virtuelle numérique qui est munie d'une capture de mouvement. La qualité très élevée d'effet visuel 3D que l'on pouvait atteindre dans des systèmes traditionnels VFX ne pouvait absolument pas être recréée de façon identique avec le système de production virtuelle. Principalement, à cause des enjeux de latence car pour faire performer l'environnement 3D en temps réel en gardant l'interconnection sous faible latence de toutes les composantes de production virtuelle, il fallait restreindre les objets numériques, leurs dimensions, leurs poids, leurs résolutions et les effets en temps réels dans la scène.

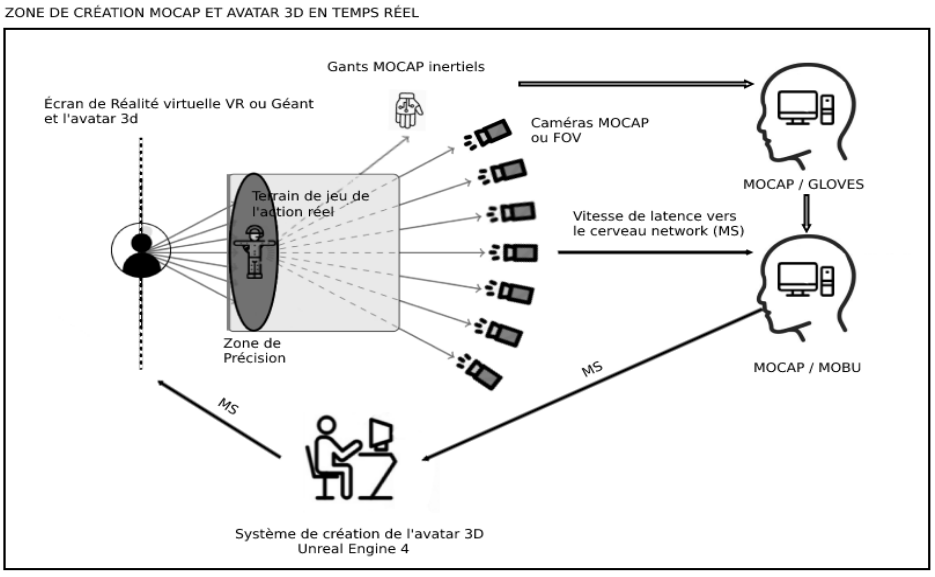

Figure 2 : Diagramme de la création d'avatar 3D avec MOCAP traditionnel en temps réel. © David Hurtubise

Pour atteindre une telle fréquence d'images, le système de capture de mouvement en temps réel devait utiliser deux ensembles de données :

- Les données en temps réel qui sont transmises au réseau à partir de capteurs MOCAP sur le corps des utilisateurs, des caméras ou des décors.
- Les données prédictives qui sont calculées par le système programmé dans Unreal Engine. Qui « remplissent les espaces vides » pour reproduire les données en direct d'un utilisateur afin de contrôler la latence et la fréquence.

L'ajout de modèles de programmation vient pallier le manque de précision entre données et latences lors de la synthèse finale de l'image.

3.3 RÉCIT DE PRATIQUE 3 – LE MOCAP AVEC IA

Comme il s'agit de la principale proposition de ce mémoire, j'invite le lecteur à prendre connaissance des deux itérations de Tensor Flow en annexe. La démonstration sera utile pour mieux comprendre l'argumentaire des prochaines lignes.

Pour mieux comprendre la forme d'animation dans l'espace avec une solution d'IA pour MOCAP, nous avons fait un prototype et une première phase de recherche avec un collègue d'Epic Games, Alban Bergeret, en 2019.

Le projet était basé sur un processus de conception itératif et utilisait Deep Learning avec la solution Tensor Flow. À partir des performances de l'utilisateur capturées par webcam, nous pouvons prédire les poses corporelles correspondantes d'une manière qui exploite la cohérence temporelle du mouvement humain. Le code open source du projet FaceFlow est disponible également en annexe.

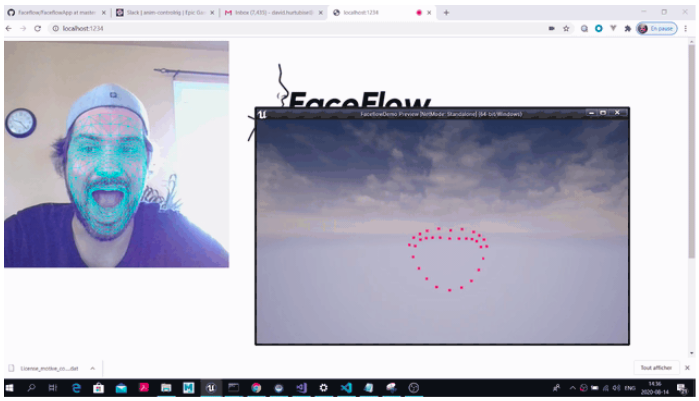

Figure 3 : Prototype itératif sur l'analyse par "Deep Learning" avec Tensor Flow dans UE4.

David Hurtubise, Alban Bergeret

En appelant uniquement la dépendance du modèle via JSON, REST API et NODE.JS, nous avons pu facilement avoir une application Web qui utilise la webcam de l'ordinateur de l'hôte. L'automatisation convoitée de l'utilisation de l'IA pour MOCAP ajoute de la complexité à l'étape d'étalonnage. Parce qu'il n'y a précisément plus besoin d'étalonnage avec FaceFlow. Votre pose faciale 3D dans Faceflow, par exemple, est automatiquement générée par les étapes dites de prédiction des modèles.

Donc, suivant la recherche de L. Yuan en 2017 (A convolutional neural network based on TensorFlow for face recognition), on utilise le modèle FACEMESH de tensorflow pour créer une application. Cette application va démarrer un serveur et un client qui va générer les informations de la webcam et les renvoyer au serveur web qui peut être soit directement celui de tensorflow ou un serveur web local. L'application Faceflow et le tracking 3D sont finalement présentés de manière simple en forme de page web. En simultané, le serveur relais des informations par Plugin sur un projet Unreal Engine et les points de la bouche sont diffusés en temps réel dans l'engin de jeu 3D.

La résolution de problème avec ce modèle possède de nombreux avantages, tel qu'une capture de mouvement simple avec n'importe quelle source vidéo et sur n'importe quelle

plateforme. Rappelons qu'à ce moment de la recherche, il n'existe que la solution avec Iphone pour la capture faciale avec Unreal Engine qui est simplifié pour n'importe quel utilisateur. Tout autre système de capture tel qu'une caméra RGB demande une programmation plus intense qui n'est pas donné à n'importe quel artiste. Aussi, le fait que la librairie de poses pour le modèle de FACEMESH existe déjà, la création de banque de mouvement n'est pas nécessaire et la reproduction de poses complexes devient alors d'une rare simplicité, et cela sans l'utilisation d'aucun capteur.

Mais les problèmes découverts lors de la création de Faceflow étaient la rigidité du système d'apprentissage et encore une fois, la latence. L'utilisation de la plupart des solutions d'IA nécessite une évaluation des images par une prédiction d'un modèle. Dans ce prototype spécifique, nous avons utilisé un modèle pré-entraîné appelé FACEMESH, qui est l'un des modèles disponibles dans la solution Tensor Flow. Cela signifie que nous devons utiliser un modèle établi à partir d'une base générée d'images qui ne sont pas les nôtres et qui ne peuvent pas être modifiées. Autant que cela puisse être la force de l'application Faceflow, autant cela nuit à la création d'autres poses et alors au perfectionnement d'un système qui se rapprocherait de l'œil synthétique et de sa précision. Sans oublier que le système ne peut dépasser une vitesse d'exécution de 15 FPS dans le serveur, ce qui, avec les renvois de données vers d'autres clients tel que dans Unreal Engine, donnait une très mauvaise latence combinée au final.

Un autre problème était la programmation de base du logiciel qui était fait en programmation web, soit JSON, NODE.JS. Ce choix de programmation aide à simplifier la création de ports entre serveurs et clients et également à créer des pages web simples. Mais il reste que c'est un très mauvais choix de langage de code lorsque l'on parle de latence.

3.4 RÉCIT DE PRATIQUE 4 – LE MODÈLE D'APPRENTISSAGE AVEC IA

En 2022, j'ai rencontré Caroline Louet qui travaillait sur une application pour la langue des signes et qui cherchait à avoir plus de liberté de création pour créer une série animée 3D pour enfants sourds et muets. Avec l'aide de Caroline, nous avons créé un prototype appelé LSQ-SENS, qui est un projet open source basé sur le travail de Nicholas Renotte et sa démo complète de détection de la langue des signes en utilisant Python. Soit le Modèle d'apprentissage en profondeur LSTM.

Nous avons créé notre démo dans le contexte de la langue des signes québécoise. Cette démo open source exploite un modèle de détection de poses pour créer une séquence de points clés qui peut ensuite être transmise à un modèle de détection d'action réels.

Il peut être défini en **3 étapes** :

Étape 1 :

Nous collectons des données d'image à partir de points clés de poses de Mediapipe et les enregistrons en tant que collection numpy.

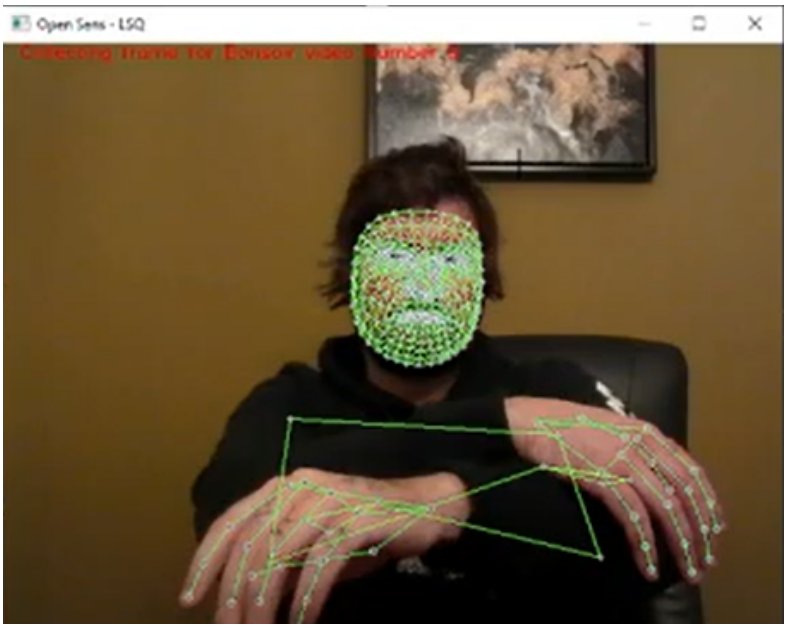

Figure 4 : Prototype itératif de LSQ-SENS. David Hurtubise, Caroline Louet

L'extraction de pose est un processus DL qui résout l'un des problèmes de la capture de mouvement traditionnelle qui est l'occlusion.

Étape 2 :

Nous formons un réseau de neurones profonds avec des couches de mémoire à long terme. Nous pouvons donc extraire l'action pour un certain nombre d'images au lieu d'une seule image. Cela se fait à l'aide de la solution Tensor Flow. Et il crée un modèle personnalisé au lieu d'utiliser le modèle préformé de TS.

Étape 3 :

Nous combinons avec OpenCV les étapes 1 et 2 pour prédire en temps réel l'action à l'aide d'une caméra RVB, dans cet exemple une webcam. OpenCV donne la possibilité d'appeler n'importe quel type de caméra dans le processus.

Figure 5 : Étape 3, Prédiction en temps réel avec un modèle DL. David Hurtubise, Caroline Louet

Le code et le modèle DL personnalisé de LSQ-Sens sont disponibles en annexe. Cela a résolu le problème de rigidité du système et a donné plus de contrôle sur la création et le développement de nouveaux concepts d'étalonnage pour MOCAP.

Il reste que le temps propice à la création des EPOC, soit la création des apprentissages est un processus extrêmement coûteux en temps et en consommation GPU afin d'affiner les données rassemblées par webcam. Nous pouvons appeler ici un code à génération lente qui demande une courbe d'apprentissage importante de l'ordinateur mais aussi de l'humain qui doit affiner et comprendre toute cette programmation.

C'est en été 2022, durant mon travail chez Rodéo FX que j'ai vraiment commencé à regarder le projet de *Deepfake* avec DeepfaceLive. Je m'intéressais depuis un moment aux modèles génératifs GANs, et la suite logique de la recherche était de tester également la création

d’une animation 3D avec capture de mouvement qui serait par la suite optimisée par AI avec un procédé d'encodage automatique de l’image.

J’ai donc installé DeepfaceLive sur mon laptop et créé un avatar numérique avec l’outil de Méta Humain de Epic Games. Il est possible de faire l’apprentissage automatique de n’importe quel modèle pour l’encodage final avec DeepfaceLive. Donc, un peu à la manière de LSQ-Sens, j’aurais pu refaire mon propre visage en suivant une série d’étapes à l’apprentissage des EPOC. Un procédé qui prend une bonne semaine de temps en puissance GPU pour la création d’un modèle qui est utilisable. Il est essentiel d’avoir une bonne collection d’images de visages, soit facilement entre 7 0000 et 10 000 images pour pouvoir encoder toutes les expressions et les poses. Ce qui est beaucoup plus exigeant que la série de 54 poses pour des modèles 3D comme Faceflow. Les faces 2D prises par caméra sont introduites dans un modèle et traitées par GPU, puis sont convertis en information vectorielles. Par la suite, suivant une analyse de réseau générative adverse, deux entités sont formées dans un encodeur partagé, et sont toujours mis en relation afin de sortir la meilleure version pour l’encodeur.

Par manque de temps et de puissance GPU pour me recréer en modèle d’apprentissage, j’ai choisi d’utiliser l’un des modèles déjà en librairie de DeepfaceLive, soit le Joker. En créant un avatar Méta humain 3D qui avait les cheveux et la forme de tête du Joker, sans avoir toutefois tous les traits spécifiques au visage, j’ai pu le combiner en temps réel dans DeepfaceLive. Il est d’ailleurs possible de voir la vidéo en annexe.

Figure 6 : Prédiction en temps réel avec un modèle DeepFaceLive et Meta Human. David Hurtubise

Ce procédé était le meilleur résultat en termes de reproductivité de la qualité de l'animation et de l'image que j'ai réussi à avoir de toutes les itérations dans ce mémoire. Une reproduction du visage du Joker avec un simple Méta humain en référence était rapidement réalisable. De plus, l'utilisation d'une source vidéo permet l'utilisation de Méta Humain dans Unreal Engine directement, et donc d'utiliser un Iphone pour la capture de mouvement en temps réel. Il est donc clair que l'utilisation de base de données d'apprentissage automatique est la clé vers une simulation machine quasi similaire avec l'œil humain.

Mais, la capture de mouvement de l'IA par analyse de modèles avec AI Learning est limitée à la définition principale du dictionnaire de l'apprentissage : Préparer, Présenter, Pratiquer, Performance. Il s'agit de créer et d'utiliser un modèle pour faire une prédiction. Le Deep Learning est plus lent à traiter les informations que la capture directe de mouvement avec des capteurs. Le traitement est également limité par les forces des ordinateurs en réseau qui restituent l'information. Ce qui fait que pour une capture de mouvement utilisant DeepFaceLive et un MetaHumain, nous parlons d'une énorme latence avec une vitesse de 4 fps après tous les procédés de calculs des différentes composantes.

Peut-être que le futur proche permettra la création de mouvements d'IA ultra-rapides simplement par randomisation, mais il reste que tout système de randomisation a besoin par logique d'une définition de base linéaire et claire. Ce qui amène la résultante ; la latence et l'imprécision des solutions d'IA. Par exemple, les modèles ML de Tensor Flow sans les éléments d'encodage comme *DeepFake* ont une latence limitée à 15 fps. Quant aux systèmes de capture traditionnels comme Optitrack, on peut monter jusqu'à 250 fps.

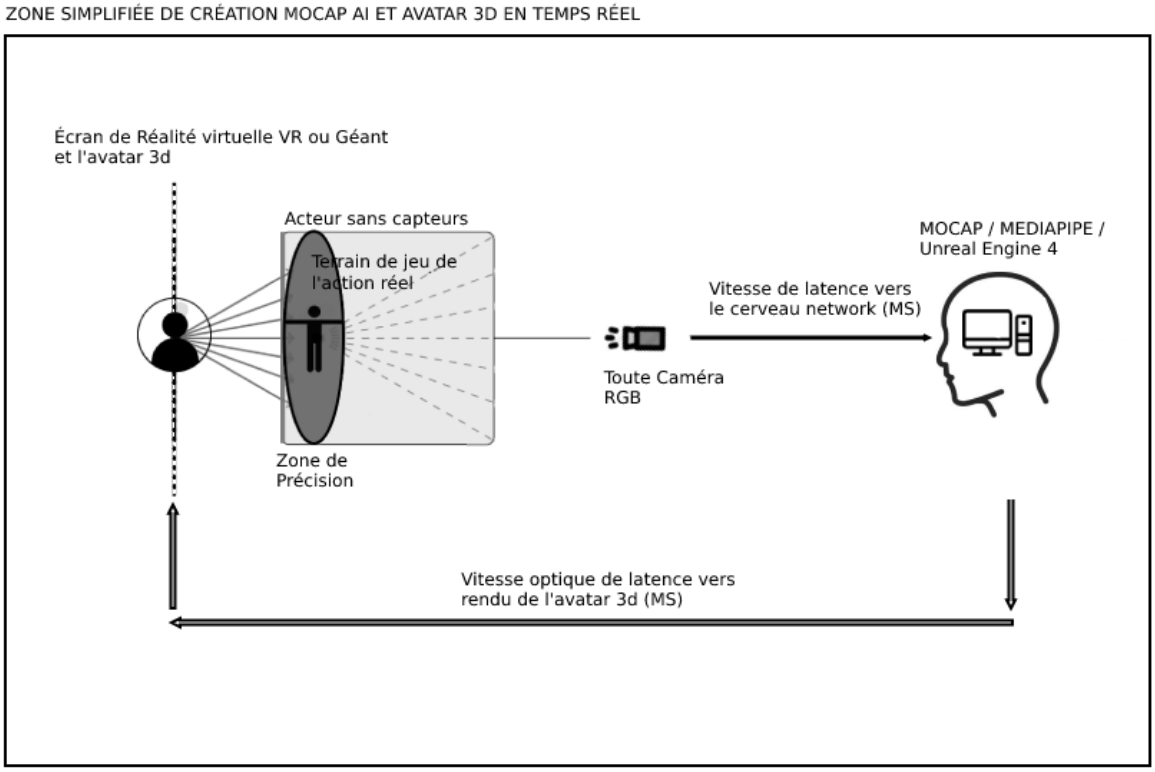

Figure 7 : Diagramme de la création d'avatar 3D avec ML en temps réel. © David Hurtubise

CHAPITRE 4

ANALYSE DES RÉSULTATS

Il faut comprendre que notre proposition repose principalement sur la conception d'un système d'échange. Chaque flux de données influence l'expérience globale selon les paramètres définis. Un changement sur une variable influence l'ensemble soit par une augmentation ou une diminution de la performance des autres variables. La performance du système ne repose pas sur la quantité d'informations mais bien sur son optimisation et la séparation des tâches de traitement.

4.1 PROPOSITION DU SYSTÈME IA OPEN SOURCE POUR MOCAP

Mise en garde : l'IA est un secteur immense qui touche une foule de nouveaux secteurs. Le Machine Learning a également des incidences sur de nombreux secteurs du Computer Vision. L'objet de ce mémoire se situe plutôt du côté d'une analyse globale des paramètres. Et comment l'analyse de ceux-ci permettent de mieux comprendre comment un ordinateur interprète l'environnement réel à partir de ses systèmes de calculs.

Il faut comprendre que notre proposition repose principalement sur la conception d'un système d'échange. Chaque flux de données influence l'expérience globale. Un changement dans une variable influence l'ensemble en augmentant ou diminuant la performance des autres variables. La performance du système ne repose pas sur la quantité d'informations mais sur son optimisation et la séparation des tâches de traitement. Bref, les données fournies dans ce mémoire sont interdépendantes les unes des autres afin de reproduire la complexité de la copropriété entre l'hôte et l'avatar 3D.

Pour ce mémoire, nous avons défini de travailler les paramètres de calcul pour l'optimisation des avatars 3D en temps réel avec différents systèmes de capture de mouvement tel que :

- La Latence
- Les Paramètres d'apprentissage
- Le Reciblage en temps réel

Tableau 2 : Regroupement des paramètres de MOCAP.

Solutions de MOCAP	Latence	Paramètres d'apprentissage	Reciblage en temps réel
Optitrack	240 FPS	Modèles ROMs. Motive	Calibrage avec une série de poses. Besoin de logiciels tiers pour l'animation 3D. Précision excellente. Qualité de rendu difficile à obtenir.
FaceFlow	15 FPS	Modèles préformés DL. FaceMesh	Pas d'étalonnage. Limité aux infos du modèle préformé. Précision Moyenne. Qualité de rendu faible.
DeepFake	4 FPS	Modèles Open Source DL. DeepFaceLive	Offre la possibilité de créer de nouvelles formes et étapes d'étalonnage. Précision Moyenne. Qualité de rendu excellente.

Voici une analyse système très comparable à la thèse de recherche de Soumil Chugh (An Eye Tracking System for a Virtual Reality Headset, 2020). S. Chugh qui était basé sur les paramètres de solutions de suivi VR bien connues telles que Tobii, Pupil Labs et SMI.

Suivant la logique de ce tableau, toute diminution des paramètres définis dans cette note, apporte également une modification négative de la précision de la mesure de l'erreur entre la capture réelle et virtuelle, comme le montrent les résultats du tableau 2.

Nous nous attendons normalement à ce que le système Optitrack offre une grande précision de reconstruction globale du corps et offre également une latence relativement faible. Cependant, ce n'est pas le cas, en raison du temps de préparation élevé et du caractère invasif du système. Ensuite, le système Oculus pourrait avoir l'avantage d'une grande précision et d'une faible latence. Cependant, il peut altérer le sens de la corrélation entre le monde réel et virtuel par des estimations incorrectes des positions en raison de sa limitation à l'étalonnage des ROM. Et si nous prenons la solution Tensor Flow, l'IA devrait optimiser l'efficacité de la reconstruction car elle supprime les étapes d'étalonnage, mais les fps limités et les modèles de pré-entraînement limités en font une solution très rigide et bon marché. Cependant, le prototype final utilisant le modèle Open source DL offre les plus grandes possibilités d'hybridation avec une qualité excellente de rendu avec *DeepFake*, mais se limite à 4 fps après avoir suivi toutes les étapes d'encodage en temps réel car cela demande un procédé faramineusement coûteux en GPU.

Chacun des récits de pratiques réalisés dans cette phase de recherche création apporte un un soutien à la compréhension de la capture de mouvement et du reciblage d'avatar 3D par ordinateur et permet de regrouper une étude plus détaillée des paramètres de MOCAP. Si nous revenons au récit de pratique 1 avec le pipeline de Laflaque, celui-ci permettait de mieux comprendre les étapes et la structure logiciel / Hardware pour la création d'un avatar, donc la reproduction de l'image 3D par ordinateur. Soit les étapes des paramètres d'apprentissages. Mais de nombreux défis de paramètres qui apportent eux-même des défis de création ont apporté des ajustements au diagramme du MOCAP traditionnel en lien avec la latence du système et le temps plus lent de reciblage en temps réel. Il s'agissait d'un procédé hybride entre traditionnel et temps réel qui ne permettait pas de reproduire directement un schéma de l'œil synthétique. Un récit de pratique qui ne permettait pas d'accentuer l'état d'Embodiment d'un artiste vers un avatar et donc de reproduire une physicalité propre au schéma de l'œil humain.

Ce qui nous emmène au récit de pratique 2, sur la production virtuelle. Ce procédé qui est en temps réel et sous très faible latence, permet un rapprochement entre la fonctionnalité et la

schématisation de l'œil vers l'ordinateur en temps réel. La programmation des éléments de MOCAP pour le déploiement visuel sur le mur LED ajoute à la rapidité du calcul GPU. Le système qui implique encore une structure logiciel / hardware mais qui est accentué par une programmation logicielle avec Unreal Engine, permet de simuler de plus en plus la latence précise de l'œil humain avec un ordinateur. Il reste par contre que dans le récit de pratique 2, un bon nombre de calibration et de paramètres d'apprentissages doivent être effectués de manière manuelle. La calibration régulière du système Optitrack pour pouvoir garder une précision de capture en est un bon exemple.

Afin de pallier au manque d'autonomie des 2 premiers récits de pratique, vient le récit de pratique 3 et 4, qui permettent d'évaluer l'intégration de système AI dans la structure d'une production virtuelle. Nous avons d'ailleurs rapidement découvert dans ces 2 pratiques que le AI est effectivement une excellente solution et que l'apprentissage automatique donne un apport fulgurant au paramètre d'apprentissage du MOCAP.

Nous pouvons affirmer que l'utilisation de l'apprentissage automatique avec IA, principalement avec les solutions *DeepFake*, apporte de nouvelles informations à la question de la recherche: Pour le reciblage en temps réel, comment rapprocher le plus possible les yeux synthétiques de l'ordinateur des mécanismes de l'œil humain ?

CONCLUSION

En résumé, les récits de pratiques ont permis de constater la limite des systèmes traditionnels pour affronter la complexité des types de captures de mouvement. En suivant l'approche itérative du présent mémoire, la suite logique de la recherche s'est orientée vers le développement de solutions d'automatisation des paramètres de traitement des données de captures à l'aide des technologies AI.

La conclusion est donc la suivante : cette recherche nous a permis de bien définir les paramètres et facteurs de la MOCAP en général, mais ne répond pas de manière complète à la question de recherche qui est: Pour le reciblage en temps réel, comment rapprocher le plus possible les yeux synthétiques de l'ordinateur des mécanismes de l'oeil humain ? La réponse à la question posée n'est donc plus comment l'ordinateur voit le monde mais comment il permet au spectateur de se voir autrement que selon sa propre vision humaine.

Par contre, la volonté de mieux saisir la logique inhérente aux systèmes à motivé de multiples expérimentations et recherches. Régler les problèmes de latence permet de se rapprocher des mécanismes de l'œil humain et travailler sur l'apprentissage machine a pour principale conséquence d'imaginer des systèmes sans capteurs et adaptés à de multiples corps et types de mouvement. En ce sens, même si cet aspect n'est pas évident à prime abord, le fait d'utiliser une simple caméra d'ordinateur pour la capture de mouvement offre un immense potentiel pour la création, comme nous l'avons souligné au tout début de ce mémoire, d'avatars au sein de métavers.

Ce qui amène à comprendre que malgré les désavantages drastiques de l'A.I. et du machine learning sur la latence, celle-ci devrait être approfondie pour permettre un nouveau récit de pratique entre la production virtuelle et le MOCAP par AI. Une nouvelle hypothèse sur une optimisation qui fonctionnerait sans l'usage de capteurs, qui diminuerait le temps de calibration des

systèmes mocap et qui permettrait une forte automatisation du rendu. En somme, par la transparence de l'interface, dans une perspective artistique, l'usage de l'IA permettrait d'intégrer la capture de mouvement dans une foule de nouveaux secteurs du divertissement. Dans le contexte de l'usage des métavers, il semble logique que ces solutions s'installent puisqu'elles deviennent accessibles au plus grand nombre, dans une logique disruptive et automatique, de technologies avancées de capture.

Il reste encore beaucoup de travail avant de mettre en place une architecture de production virtuelle rapide et efficace qui définit le mieux le sens de la vision par ordinateur. Ou les yeux humains. Et les yeux, en termes de physiologie sont stéréoscopiques, chaque œil obtient une image légèrement différente, en plus d'un traitement très rapide et imaginatif comme le cerveau, a une meilleure procédure d'étalonnage pour réduire l'ambiguïté dans la reconstruction des voxels et donc fournir de meilleures données de voxels. Actuellement, le processus de modélisation, de suivi et de combinaison de l'objet réel dans un modèle 3D complet se fait en plusieurs étapes et de façon très linéaire. Mais si nous utilisons un système d'IA d'apprentissage automatique, ce sera principalement de manière à faire évoluer l’image d’une façon beaucoup plus neuronale, non linéaire et générative.

Une prochaine recherche serait de connecter ces informations de recherche à un système de production virtuelle. Soit la connectivité entre la production virtuelle et le AI en *DeepFake*. Mais en gardant en tête de toujours contrôler un niveau de latence acceptable.

BIBLIOGRAPHIE

Peter Jarvis. (1987). *Adult learning in the social context*, Londres, Croom Helm.

Liping Yuan. (2017). A convolutional neural network based on TensorFlow for face recognition, IEE.

David Hurtubise (2021). Design Experientiel avec Tensorflow, VR et *Embodiment*. Repéré à MUTEK
https://www.youtube.com/watch?v=YeHHNQT3UC4

Daniel Kolb. (1984). Experiential learning Theory (ELT), Publié dans *PrenticeHall.*

Stévance, S., & Lacasse, S. (2014). Les enjeux de la recherche-création en musique. *Presses de l'Université Laval.*

Epic Games. (2019). Production virtuelle dans UE4 | SIGGRAPH 2019 Unreal Engine. Repéré à https://www.youtube.com/watch?v=rbnXa0SVKT8

Laban. (1966). A synthesized act of perception and function, 7.

David Hurtubise & Caroline Louet. (2022). Récit de Pratique 4: LSQ - Senses. Repéré à https://www.youtube.com/watch?v=5c1rc4XhNiQ

Albert Menache. (2000). Understanding Motion Capture for Computer Animation and Video Games. San Diego/San Francisco/New York/Boston/ London/Sydney/Tokyo: *Morgan Kaufmann/Academic Press.*

Epic Games. (2018). Mixed Reality with UE4 | SIGGRAPH 2018 Unreal Engine. Repéré à https://www.youtube.com/watch?v=4RHb1q2Ng3A&t=91s

Louis-Claude Paquin. Cynthia Noury. (2018), Définir la recherche-création ou cartographier ses pratiques? ACFAS. Université du Québec à Montréal. Dossier: Recherche-Création

Barry J. Purvis. (2014), Stop-Motion Animation: Frame by Frame Film-Making with Puppets and Models.*Bloomsbury Publishing.*

Epic Games. (2018). ICI Laflaque - Broadcast quality within a tight timeline. Repéré à https://cdn2.unrealengine.com/Unreal+Engine%2Fresources%2FLaflaque_whitepaper11.16.18-e0d9e5c363e71dac83fc4439b89b9cf2c77d4d24.pdf

Facebook. (2020). Research: Photorealistic Avatars & Full Body Tracking, VR Trailer. Repéré à https://www.youtube.com/watch?v=Q-gse_hFkJM&feature=youtu.be

Kevin He. (2018). Using Deep Learning to Create Interactive Actors for VR. DeepMotion. Repéré à Oculus Connect 5 | Using Deep Learning to Create Interactive Actors for VR.

Yinghao Huang et al. (2018). Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time". SIGGRAPH Asia 2018 Technical Pa-pers. ACM., 185.

Hamdi Mortan, N. (2014). Using game engines in interactive Co-Design. University of Cincinnati.

Fairuz Shiratuddin, M., Thabet W. (2001). Virtual Office Walkthrough Using a 3D Game Engine. Virginia Polytechnic Institute and state University.

Jacobson, L. (1991, 08). Virtual reality: a status report. AI Expert, 6(8), 26+. Repéré à http://link.galegroup.com.sbiproxy.uqac.ca/apps/doc/A11056777/CDB?u=crepuq_uqac&sid=CDB&xid=a21b3038

Epic Games. (2019). THE VIRTUAL PRODUCTION FIELD GUIDE VOLUME 1. https://cdn2.unrealengine.com/vp-fieldguide-v1-3-01-f0bce45b6319.pdf

Metaphisic.ai (2022). The Future of Autoencoder-Based Deepfakes. Repéré à https://metaphysic.ai/future-autoencoder-deepfakes/

Prabu. (2020). The Virtual Production of The Mandalorian, Season One. Repéré à https://www.vfxexpress.com/thevirtualproduction-of-the-mandalorian-season-one/

A. (2020). New Virtual Technologies Remake VFX'S Future Pipeline. Repéré à https://www.vfxvoice.com/new-virtualtechnologies-remake-vfxs-future-pipeline/

Giannalberto Bendazzi. (1994). Cartoons: One Hundred Years of Cinema Animation. Bloomington, *University of Indiana Press.*

ANNEXE 1

EXTRAITS DE JOURNAL DE RECHERCHE

(2019) MUTEK, Design Expérientiel avec Tensor Flow, VR et Embodiment.

https://www.youtube.com/watch?v=YeHHNQT3UC4&feature=youtu.be

(2020) FaceFlow

https://github.com/Hurtubise-David/Faceflow

(2022) Recherche-Création et art numériques: autopsie des pipelines de production.

https://github.com/Hurtubise-David/openmetaverse/tree/main/projects/LSQ-Sens

(2022) DeepFake avec Méta Humain.

https://www.youtube.com/watch?v=fjyorBPavJg