Modeling of reactive movement and non-verbal behaviors for the creation of digital agents for virtual reality.

Alberto Jovane

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *informatique*

Par
Alberto JOVANE

Modélisation de mouvements réactifs et comportements non verbaux pour la création d'acteurs digitaux pour la réalité virtuelle

Thèse présentée et soutenue à Rennes, le 27 - 02 - 2023
Unité de recherche : Irisa/Inria

Rapporteurs avant soutenance :

Rachel McDonnell	Maîtresse de Conférences, Trinity College, Dublin
Michael Neff	Professeur, Université de Californie

Composition du Jury :

Président :	Rachel McDonnell	Maîtresse de Conférences, Trinity College, Dublin (rapporteur)
Examinateurs :	Nuria Pelechano	Enseignante Chercheuse á Universitat Politecnica de Catalunya
	Sylvie Gibet	Professeure, Université de Bretagne Sud
Dir. de thèse :	Julien Pettré	Directeur de Recherche à Inria

Invité(s) :

Ludovic Hoyet	Chargé de Recherche à Inria (encadrant)
Claudio Pacchierotti	Chargé de recherche au CNRS (encadrant)
Marc Christie	Enseignant Chercheur, Université de Rennes (encadrant)

Acknowledgement

By reading these pages, you will find just a tiny scientific contribution in the field of digital humans, developed over three years (and a bit more) of PhD research. But, unfortunately, you will not be able to see what lies behind these pages. Each thesis holds a backstory, shaped by the human environment in which each student operates. When I think of this thesis, I like to remember a portion of time when the lives of various individuals intersected. Each of these persons made a unique contribution to this experience by participating in long working sessions, meetings, coffee breaks, but also parties, evenings out, sports sessions, brief chats in the corridor, or online chats and video calls (which were particularly significant during covid time).

I start by thanking my first team, Mimetic, which welcomed me at the beginning of my master's stage and shared time with me throughout the entire PhD. Thanks to my first supervisor, Marc, who introduced me to the world of research, showed unexpected trust in my early ideas, gave me so much confidence, and continued to supervise my work during the PhD. Thanks to Nathalie, who assist me throughout my entire time in Rennes, along with Hélène, and Gwenaelle. To my first office mates: Anthony, Ludo, and Gwendal, and all the gang: Rebecca, Benjamin, Nils, Amaury, Diane, Arzhelenn, Olfa, Xi (and all the rest of the "mimebrid" group), the visiting students Sheryl and Robyn, and the others who shared their time with us, thank you all, you really made me feel at home even if I couldn't speak a single word of French, and you helped me so much. I am not able to imagine a better group of people to meet when moving to a new country.

When I started my PhD, I met the rest of my official supervisors: Julien, Anne-Helene, Ludovic, and Claudio, along with Marc, as well as my "additional" and very important supervisors Katja and Pierre. I can't thank all of you enough for your wise and constant support. There was always someone available to help me, and without you, I would not be here writing these words. It is also important to thank the European commission which founded this project (PRESENT) and all the partners who participated and shared their knowledge with us. I also want to thank Adele, my office mate and companion in this

European project. Thank you for all the support; it has been great to have you share the same experience.
During the first year of my PhD, when I joined the Rainbow team, I had the chance to meet another group of incredible people. I start by thanking the group of Italians: Claudio, the first two Marcos (Aggravi and Cognetti), Alexander (source of inspiration for visual servoing) along with the important addition of Adam (who was also a great flatmate), Naty, Chiara, Gaetano, Riccardo, more recently Nicola, the third Marco, Antonio, and the boss Paolo. Next, I would like to thank all the rest of the Rainbow team, including the early members Julien, Fabian, Florian, Wouter, Rahaf, Ramana, Joudy as well as the newer members Elodie, John, Lev, the two Maxime, Lisheng, Erwan, Thibault, Pascal and Samuel. Thank you, it has been great to spend time with all of you.
During the second half of my PhD I joined the newly established VirtUs team. I would like to thank the people with whom I share this significant phase of my PhD: Thomas (my last office mate alongside Adele), Tairan (my previous office mate), Gwenaelle, Yuliya, Solenne, Nena, Robin, Alexis, Philippe, Remi and many others. I need to extend my thanks to people from various team, or external or visiting who contributed to making my time in Rennes great: Stephane, Luise, Axel, Marie, Salomè, Hasnaa, Abdul, Raul, Ruben, Lucia and more. Additionally, a big thanks to the wonderful people working at the cafeteria, you made each day better. I also need to give a special thanks to Federico and Vernoica (my first great flatmates here in Rennes), Max, Octavie, Monik, Lorenzo, the basketball group at INSA, the football team, Khadi and Philippe (for the great support and all the running activity). I would also need to thank Filippe Marlij, for the artistic support, and Richard Benson, for being a source of inspiration. A big gracias to my distant friends from Barcelona (Emeline, who also provide a valuable scientific support, Alex, Fabio, Jerome, Kymry, Mathieu, Micheal, Polli, Oscar, Juan), all my friends in Italy (non posso nominarvi tutti qui, tanto sapete chi siete), and the team of screenculture film festival. A special merci to Fanny for her unique support. And finally a special thanks to my family (madre, padre e Dario), grazie davvero per tutto quello che fate sempre. Grazie al resto della famiglia: ai cugini (Vincenzo, Paolo, Domenico, Nino, Rossana, Daniela, Paolo, Valeria, Fabio, ...), agli zii e ai nonni sia presenti che non più, sia in Italia che in Brasile, e anche un Grazie a quelli che anche senza essere familiari diretti considero come parte della famiglia, grazie di tutto.
An additional thanks to the commission and the reviewers of this manuscript which spend time to evaluate it.

Please forgive me if I forget your name, or I mention it and misspelled it, or if I put you in a wrong team. Each one of you is important to me, thank you all.

Résumé en français

Au cours des cinquante dernières années, les technologies numériques sont devenues un élément stable de notre vie quotidienne. Ces technologies offrent diverses fonctionnalités: aide aux tâches quotidiennes, accès ou création de contenus divertissants, systèmes de communication et bien d'autres encore.

Cette popularité soulève la nécessité de réduire l'écart entre le numérique et le réel, en termes de représentation, d'expérience et d'accessibilité. Cela signifie que les technologies doivent évoluer pour générer un contenu numérique tridimensionnel photoréaliste pour les films et les jeux vidéo, pour développer des moyens d'accéder au "monde virtuel" par le biais d'appareils plus immersifs, pour proposer des interactions plus humaines avec le contenu numérique, *etc.* En ce sens, nous avons tendance à abstraire et à repenser constamment la communication homme-machine pour accroître l'accessibilité et la facilité d'utilisation, par exemple en donnant une voix à nos appareils, comme dans le cas de la navigation GPS (Global Positioning System) qui utilise la voix humaine comme guide, et plus récemment la simulation d'assistants vocaux sensibles capables d'interpréter nos questions et d'y répondre [Seaborn et al., 2021]. En regardant encore plus loin, nous imaginons un monde où davantage d'interactions auraient lieu dans un "meta-world" (un espace social virtuel reliant le monde entier), y compris des interactions multimodales complexes avec des utilisateurs distants ainsi qu'avec des personnages simulés. Dans ce contexte, la génération d'humains virtuels fidèles et réalistes est fondamentale. Des domaines tels que le divertissement (films, jeux vidéo, télévision, *etc.*), les simulations et les assistants numériques s'appuient déjà sur des reproductions simulées d'humains virtuels (voir la figure 1).

Cette thèse a été créée dans le cadre d'un projet de recherche européen plus large appelé PRESENT [1], dont l'objectif était de créer des humains virtuels sensibles et très réalistes. L'objectif spécifique de cette thèse est lié à la simulation et à l'interprétation du mouvement de ces humains virtuels sensibles pour transmettre des informations non verbales.

1. https://www.upf.edu/web/present

a b c

Figure 1 – Quelques exemples d'applications utilisant des humains virtuels développés dans le cadre du projet PRESENT. L'exemple (a) montre une scène d'une émission sportive dans laquelle le présentateur (à gauche) est assisté par un agent virtuel (à droite), développé par Brainstorm [2]. L'exemple (b) montre un extrait d'une expérience artistique en réalité virtuelle développée par CREW [3]. L'exemple (c) montre le totem numérique utilisé pour l'interaction avec EVA (assistant virtuel ETIC), développé par l'université Pompeu Fabra de Barcelone [4] pour accueillir et guider les invités et les étudiants de l'université.

Nos objectifs de recherche explorent divers aspects de ces interactions entre l'homme et l'homme virtuel. Nous définissons l'objectif de simuler des réactions crédibles, en termes de mouvement du corps et du regard, pour les agents virtuels interagissant avec les utilisateurs. Pour ce faire, notre intuition est d'adapter ces caractéristiques de la dynamique des humains virtuels en fonction de la façon dont le mouvement apparaît du point de vue de l'utilisateur. En effet, lors de nos interactions quotidiennes, nous essayons d'influencer la façon dont les autres nous perçoivent, en utilisant des éléments tels que notre distance ou notre proximité, la direction de notre regard, mais aussi notre posture, la vitesse ou l'amplitude de nos gestes, *etc.*Nous adaptons souvent ces indices en fonction d'un objectif précis. Nous adaptons souvent ces indices à un ou plusieurs interlocuteurs, afin de communiquer nos intentions et nos émotions. En outre, les indices de mouvement sont également l'expression de caractéristiques plus implicites de la communication, telles que les traits de personnalité, la fatigue, le contexte social, *etc.*Par conséquent, nous nous attendons à ce que les humains virtuels se comportent de la même manière dans un contexte virtuel. Cette approche définit la contribution principale de cette thèse, un système pour éditer les caractéristiques du haut du corps d'un personnage virtuel, basé sur l'apparence visuelle de ces mouvements du point de vue d'un observateur externe (*e.g.* l'utilisateur).

2. https://www.brainstorm3d.com/
3. https://crew.brussels/
4. https://www.upf.edu/

La génération de telles adaptations sur les mouvements du corps et du regard nécessite une compréhension approfondie de la façon dont ces indices sont perçus par l'utilisateur. C'est pourquoi un autre objectif de notre recherche est de mener des études perceptives sur l'adaptation des mouvements et du regard des agents virtuels. En effet, il est crucial de valider l'impact des agents virtuels sur la perception de l'utilisateur, qui dépend de plusieurs aspects, tels que l'apparence, le mouvement et le comportement de l'agent, mais aussi le moyen par lequel nous percevons l'humain virtuel (*e.g.* écran ou casque de réalité virtuelle). Par conséquent, pour approfondir l'analyse de ces interactions mutuelles, nous devons étudier les technologies que l'utilisateur utilise pour percevoir les humains virtuels dans un monde virtuel. Dans ce contexte, nous avons décidé d'axer un autre objectif sur la simulation du contact à l'aide de technologies haptiques. De telles simulations améliorent notre perception sensorielle des humains virtuels et affectent la manière dont nous interagissons avec eux. Enfin, nous considérons *staging*, *i.e.* disposition de divers éléments inclus dans un plan de caméra, pour visualiser les interactions entre les humains virtuels dans un contexte cinématographique. À cette fin, notre dernier objectif étudie les méthodologies impliquées dans la génération de caméras virtuelles afin d'optimiser l'expérience d'événements impliquant des humains virtuels dans des environnements virtuels.

Contexte

Le mouvement humain est une conjonction complexe de mécanismes multiples. D'une part, il y a la contrainte physique de la structure corporelle, composée d'os, d'articulations, de muscles, de peau et d'autres tissus, optimisée, au cours de millions d'années d'évolution, pour effectuer diverses tâches fonctionnelles (comme marcher, courir, saisir des objets, *etc.*) et pour communiquer, par ces tâches et d'autres, avec d'autres humains ou d'autres animaux, par exemple par des expressions agressives ou amicales. En outre, chaque tâche de mouvement est exécutée comme une composition synergique de multiples impulsions que nous coordonnons et apprenons à exécuter dès les premiers stades de la vie.

D'autre part, il y a l'intention, volontaire ou involontaire, et le but qui guident la génération de chaque tâche motrice. Des éléments tels que l'âge, le contexte social, la personnalité, les émotions, la fatigue et bien d'autres encore affectent l'intention et la performance. Dans ce contexte, nous pouvons affirmer que chaque mouvement peut être exécuté d'une grande variété de façons, et qu'il est difficile de reproduire exactement le même mouvement deux fois [Peng et al., 2014].

Heureusement, notre cerveau est capable de regrouper et d'interpréter la complexité intrinsèque de chaque mouvement humain perçu, en décomposant les intentions et les émotions [Mar, 2011]. Pour ces raisons, le mouvement humain est l'un des principaux moyens de **communication** interpersonnelle, mais aussi l'un des plus délicats. En effet, toute forme de communication repose sur une convention convenue entre deux parties, définie sur la base d'un ensemble de règles et d'un dictionnaire de messages, dont nous sommes plus ou moins conscients. Pour certains moyens de communication, comme les langues écrites, les règles sont plus strictes, mais l'interprétation dépend toujours de divers facteurs, comme le contexte ou le milieu culturel. Cependant, l'étude des caractéristiques du mouvement humain est pertinente dans divers domaines, depuis les études médicales et comportementales jusqu'aux arts du spectacle, tels que la danse ou le jeu d'acteur. Cette forme de communication non verbale est incluse dans la définition de la "communication non verbale".

Définition communication non verbale

"Non-verbal behavior refers to actions as distinct from speech. It thus includes facial expressions, hand and arm gestures, postures, positions, and various movements of the body or the legs and feet." [Mehrabian, 1971]

Face à un autre être humain, nous échangeons un large éventail d'informations non verbales. En effet, le message de communication est la combinaison synchrone de multiples éléments, où tout peut être pertinent et où tout ce qui s'écarte des conventions sociales peut nous alerter.

Communiquer avec des humains virtuels

Par conséquent, lorsque nous sommes face à un humain virtuel, nous nous attendons à interagir de la même manière qu'avec nos homologues réels [Hoffmann et al., 2009]. Cette attente est plus forte si l'humain a l'air réaliste, ce qui signifie que, dans ce cas, nous tolérons moins les imperfections des mouvements (ce phénomène perceptif est appelé la "uncanny valley" et a été défini par Masahiro Mori dans [1970] pour l'étude des robots humanoïdes). Depuis la génération du premier humain virtuel, proposé par l'équipe dirigée par William Alan Fetter au début des années 1960 chez Boeing industries à Seattle [Fetter,

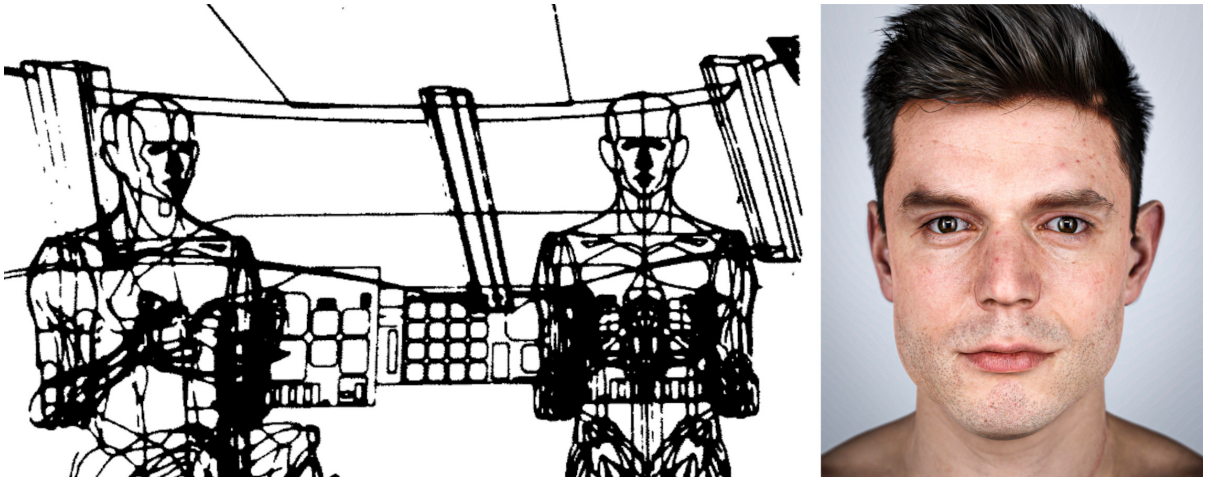


Figure 2 – Évolution de l'apparence des personnages virtuels : à gauche, le "First Man" [1968], un pilote articulé en sept segments utilisé dans le développement du Boieng 747 ; à droite, le personnage développé par Framestore [5] et CubicMotion [6] pour le projet PRESENT.

1982; Hickey et al., 1968] pour optimiser l'accessibilité des commandes d'un avion dans son cockpit (voir Figure 2), l'écart avec l'homologue réel a été réduit à la fois en termes d'apparence visuelle et de génération de mouvements.

En effet, l'emploi d'humains virtuels dans des applications interactives requiert un large éventail de compétences techniques. Principalement, comme nous l'avons dit, la synthèse d'apparences réalistes et de mouvements naturels, souvent à l'origine de l'uncanny valley s'ils ne sont pas correctement exécutés. Mais aussi, le développement de systèmes réactifs complexes pour adapter ces mouvements à l'utilisateur, comme dans les jeux vidéo où l'utilisateur contrôle le personnage virtuel (qui doit se déplacer en fonction des entrées fournies) et interagit avec d'autres personnages autonomes (qui doivent souvent se comporter et réagir de manière réaliste aux événements). Enfin, la dernière exigence consiste à fournir des systèmes et des moyens d'interaction, ainsi que la réalisation de l'environnement virtuel dans lequel les humains virtuels interagissent.

Nous pouvons identifier deux étapes principales dans l'évolution des formes d'interaction avec les humains virtuels: (i) Les premières formes d'interaction étaient unidirectionnelles, les humains virtuels apparaissant sur un écran de visualisation, c'est-à-dire le moniteur de l'ordinateur pour la simulation, dans les films et les publicités télévisées. (ii) Ensuite, avec l'évolution des technologies interactives, les humains virtuels ont eu besoin de capacités de réaction, comme pour les jeux vidéo et les interactions immersives

telles que la **réalité virtuelle**.

Définition réalité virtuelle

"Virtual Reality is a technical and scientific area making use of computer science and behavioral interfaces in order to simulate the behavior of 3D entities in a virtual world that interact in real-time among themselves and with the user in pseudo-natural immersion through sensory-motor channels." [Arnaldi et al., 2003]

En fait, la réalité virtuelle doit solliciter activement les sens pour que l'utilisateur humain ait le sentiment de faire véritablement partie du monde virtuel. Dans ce contexte, nous évaluons l'*immersion* dans un environnement virtuel en fonction des différents types d'informations sensorielles qui circulent entre l'environnement virtuel et l'utilisateur humain, et de leur similitude avec les informations réelles. Diverses technologies sont impliquées dans la simulation de ces sensations sensorielles, notamment les écrans montés sur la tête, pour le retour visuel, et les dispositifs haptiques, pour la simulation tactile.

La réalité virtuelle est souvent associée à un utilisateur actif, qui interagit directement dans le monde virtuel de son propre point de vue. Cependant, lorsque l'utilisateur n'est pas directement impliqué dans les événements, une forme d'expérience différente est nécessaire pour s'assurer que les événements lui sont présentés de manière appropriée. Dans ce contexte, la principale référence est la cinématographie. En effet, dès la conception de la cinématographie classique, une forme spécifique de communication a été développée, c'est-à-dire une grammaire visuelle avec des règles de cadrage des scènes, de transition entre les séquences – appelées cuts – et d'expression des styles visuels, afin de transmettre une narration et des émotions à un public [Bowen, 2013; Thompson and Bowen, 2009]. En tant que spectateurs, nous assimilons et partageons ces conventions, ce qui nous permet d'interpréter les intentions du narrateur. En effet, nous n'avons pas plus peur d'un train qui se dirige vers nous dans une salle de cinéma, que nous n'essayons de répondre aux salutations d'un journaliste à la télévision. En d'autres termes, nous pouvons dire que nous avons développé de nouvelles capacités à communiquer avec des schémas de communication partagés. Dans les films, le cadrage et la composition sont conçus en fonction du mouvement des acteurs, et la performance est adaptée au mouvement de la

5. https://www.framestore.com/
6. https://cubicmotion.com/

caméra. Lorsque ces règles sont adaptées à la réalisation d'environnements virtuels, on parle de **virtual cinematography**. Virtual cinematography étudie et propose des techniques pour cadrer et éclairer des environnements afin de transmettre des expériences cinématographiques d'événements, tels que les interactions d'humains virtuels.

Aperçu de la thèse et contributions

Cette thèse explore la manière de transmettre efficacement les différentes caractéristiques non verbales du mouvement humain et se concentre sur les différents aspects de l'échange de communication entre les humains réels et virtuels. En effet, chaque mouvement humain transmet des informations et si nous voulons simuler et reproduire des mouvements plus engageants et plus fidèles pour des expériences virtuelles, nous ne pouvons pas faire l'économie de l'étude des formes de communication et des aspects perceptifs du mouvement. Comme nous l'avons mentionné, il existe diverses formes d'échange d'informations tacites entre les êtres humains qui, dans les applications virtuelles, sont transmises par différents médias du monde virtuel au destinataire humain de la communication. Tout au long de ce manuscrit, nous adoptons différents termes pour définir ce destinataire, qui peut être un observateur passif, un interacteur actif, un spectateur, un public ou un participant immergé dans le monde virtuel. Le choix de la définition dépend de la forme de communication correspondante présentée dans la contribution correspondante.

Ce travail présente une exploration transversale de différentes disciplines (simulation du mouvement humain, perceptions appliquées, expériences immersives multimodales, cinématographie virtuelle, *etc.*) dans le but d'améliorer la simulation d'humains virtuels dans des applications divertissantes et interactives. Notre travail de recherche a été développé sur quatre axes de recherche.

Axe de recherche 1: *proposer des méthodologies d'animation efficaces afin de transmettre aux utilisateurs les caractéristiques non verbales des mouvements humains virtuels.* Dans ce cas, notre première contribution est une édition en ligne du mouvement du haut du corps d'un personnage virtuel pour l'adapter à des cibles visuelles définies par l'utilisateur du point de vue d'un observateur ou d'un utilisateur de la réalité virtuelle. Cela signifie, par exemple, qu'il est possible d'adapter l'orientation et l'amplitude du

mouvement d'ondulation d'un personnage en fonction de l'amplitude apparente que l'observateur perçoit depuis sa position. Si l'observateur est loin, l'agent doit agiter davantage pour augmenter sa visibilité. Pour atteindre cet objectif, nous avons introduit un nouveau paradigme qui place l'observateur du mouvement au centre, vers lequel l'animation est réalisée, et qui adapte le mouvement du personnage virtuel pour satisfaire diverses contraintes exprimées du point de vue de l'observateur.

Axe de recherche 2: *pour comprendre quels facteurs affectent la communication et comment ils sont perçus par les utilisateurs.* Dans ce contexte, nous étudions la perception des indices non verbaux de la communication. En particulier, nous présentons deux études : (i) la première étude évalue l'effet des variations de mouvement du haut du corps dans la compréhension des intentions du personnage virtuel, en relation avec le premier axe de recherche. (ii) La seconde étudie l'effet du regard dirigé et détourné d'un humain virtuel vers un participant humain, un phénomène appelé l'effet de regard dans la foule qui a été observé dans la vie réelle mais qui n'a pas été exploré dans la réalité virtuelle.

Axe de recherche 3: *étudier les technologies qui simulent le flux d'informations sensorielles entre le monde virtuel et l'utilisateur.* Dans le cas présent, nous nous concentrons sur la simulation du toucher. Nous menons une étude auprès des utilisateurs pour évaluer comment le retour sensoriel simulant le contact peut affecter la façon dont nous naviguons dans une foule virtuelle.

Axe de recherche 4: *proposer une solution technique permettant de transmettre à un public des événements en temps réel, impliquant des humains virtuels, sous la forme d'une expérience cinématographique.* Dans ce contexte, nous examinons comment représenter de manière cinématographique des mouvements humains précédemment inconnus, par l'analyse de l'environnement dans lequel ils interagissent. Nous avons conçu une technique pour placer des caméras cinématographiques et des caméras de poursuite grâce à une analyse topologique de l'environnement virtuel. Le but de ce travail est d'optimiser la génération de plans cinématographiques pour suivre le mouvement des personnages et des événements dans un environnement virtuel.

Schéma de la thèse. En suivant ces axes de recherche, nous décrivons ici la structure de ce manuscrit.
Dans le chapitre 1, nous présentons les principaux concepts et passons en revue la littérature relative aux travaux présentés. Ce premier chapitre est structuré comme suit. En particulier, nous discutons des principaux concepts liés à l'étude de la perception dans les environnements virtuels interactifs peuplés, en mettant l'accent sur les caractéristiques non verbales du mouvement et sur la façon de transmettre ces caractéristiques de manière cinématographique.
Les chapitres suivants présentent les contributions de ce travail.
Dans le chapitre 2, nous détaillons la première contribution de ce manuscrit, liée aux premier et deuxième axes de recherche. Nous démontrons cette technique à travers un ensemble de cas d'utilisation et une expérience utilisateur en réalité virtuelle (deuxième axe de recherche), afin de tester si notre technique peut aider l'utilisateur à mieux comprendre l'intention de l'agent virtuel.
Dans le chapitre 3 nous présentons la deuxième contribution de ce manuscrit, liée au deuxième axe de recherche réalisé en réalité virtuelle. Dans cette contribution, nous évaluons la présence d'un effet perceptif induit par le regard dirigé par rapport au regard détourné lors de l'observation d'une foule virtuelle. Nous présentons la configuration technique, l'étude utilisateur et une discussion des résultats
Le chapitre 4 détaille la troisième contribution, liée au quatrième axe de recherche, dans laquelle l'évaluation se concentre sur les capacités de l'utilisateur à naviguer dans une foule virtuelle. Nous étudions dans quelle mesure la simulation de contacts par le biais d'un retour haptique vibrotactile aide les utilisateurs à effectuer une telle tâche.
Dans la dernière contribution (Chapitre 5), nous passons d'un interactant interne à un observateur externe de l'interaction (quatrième axe de recherche). Dans ce chapitre, nous présentons une nouvelle approche pour positionner des caméras cinématographiques dans des environnements virtuels afin de suivre les événements, les actions et les interactions de personnages virtuels en temps réel.

Le Chapitre Conclusionconclut ce manuscrit, en présentant les remarques finales, une discussion générale du travail présenté au cours de la thèse, ainsi qu'un commentaire sur les orientations futures.

Table of Contents

Introduction

Through the last half-century, digital technologies have become a stable part of our everyday life. These technologies provide various functionalities: as help for daily tasks, for accessing or generating entertaining content, as communication systems and much more. With popularity, it raises the need of decreasing the gap between digital and real, in terms of representation, experience and accessibility. This means having technologies evolved to generate photorealistic three-dimensional digital content for movies and video games, to develop means to access the "virtual world" through more immersive apparatus, to propose more human-like interactions with digital content, *etc.* In this sense, we tend to abstract and constantly rethink human-machine communication to increase accessibility and usability, for instance by providing a voice to our devices, such as the case of GPS (Global Positioning System) navigation which uses human voice as guide, and more recently the simulation of sentient vocal assistants capable of interpreting, and answering, our questions [Seaborn et al., 2021]. Looking even further, we imagine a world where more interactions would take place in a "meta-world" (a social virtual space connecting all the world), including complex multimodal interactions with distant users as well as with simulated characters. In this context, the generation of faithful and realistic virtual humans is fundamental. Fields, such as entertainment (movies, video games, television, *etc.*) , simulations and digital assistants already rely on simulated reproductions of virtual humans (see Figure 3). This thesis was founded as a part of a wider European research project called PRESENT [7], which had the goal to create sentient, highly realistic virtual humans. The specific focus of this thesis is related to the simulation and interpretation of the motion of these sentient virtual humans to convey non-verbal information.

Our research objectives explore various aspects of such human – virtual human interactions. We define the objective of simulating believable reactions, in terms of motion of the body and gaze, for virtual agents interacting with users. To do so, our intuition is to adapt these characteristics of the dynamics of virtual humans in accordance to how the

7. https://www.upf.edu/web/present

a b c

Figure 3 – Some examples of applications using virtual humans developed in the context of the PRESENT project. The example (a) shows a scene from a sport broadcasting on which the anchorman (on the left) is supported by a virtual agent (on the right), developed by Brainstorm [8]. The example (b) displays an extract from an artistic experience in virtual reality developed by CREW [9]. The example (c) shows the digital totem uses for the interaction with EVA (ETIC virtual assistant), developed by the universitat Pompeu Fabra of Barcelona [10] to welcome and guide guests and students of the university.

motion appears from a user perspective. In fact, while interacting in everyday life, we try to affect the way others perceive us, using elements such as how far or close we position, the direction of our gaze, but also our posture, the speed or amplitude of our gestures, *etc.*We often adapt these cues toward one or more interactants, in order to communicate our intentions and emotions. In addition, motion cues are also expressions of more implicit characteristics of communication like personality traits, fatigue, social context, *etc.*Therefore, we would also expect virtual humans to behave similarly in a virtual context. This approach defines the main contribution of this thesis, a system to edit upper body characteristics of a virtual character motion, based on the visual appearance of such motions from the perspective of an external observer (*e.g.* the user).

The generation of such adaptations over body and gaze motions requires a deep understanding of how these cues are perceived by a user. For this reason, another objective of our research is to conduct perceptual studies on motion adaptation and gaze of virtual agents. In fact, it is crucial to validate how virtual agents impact the user's perception, which depends on several aspects, such as the appearance, the motion and the behavior of the agent but also the means through which we sense the virtual human (*e.g.* display or virtual reality headset). Therefore, to delve deeper into the analysis of this mutual interactions, we need to study the technologies that the user employ to sense virtual humans in

8. https://www.brainstorm3d.com/
9. https://crew.brussels/
10. https://www.upf.edu/

a virtual world. In this context, we decided to focus another objective on the simulation of contact using haptic technologies. Such simulations enhance our sensory perception of virtual human and affect the way we interact with them. Finally, we consider *staging*, *i.e.* arrangement of various elements included in a camera shot, to visualize virtual human interactions in a cinematographic context. For this purpose, our final objective studies the methodologies involved in the generation of virtual cameras to optimize the experience of events involving virtual humans in virtual environments.

Context

Human motion is a complex conjunction of multiple mechanisms. On one side there is the physical constraint of the body structure, composed of bones, joints, muscles, skin, and other tissues, optimized, during millions of years of evolution, to perform various functional tasks (such as walking, running, grabbing objects, *etc.*) and to communicate through these and others tasks towards other humans or towards other animals, *e.g.* aggressive or friendly expressions. Furthermore, each motion task is carried out as a synergic composition of multiple impulses that we coordinate and learn to perform from the early stages of life.

On the other side, there is the intent, voluntary or involuntary, and the purpose that guide the generation of each motor task. Elements, such as age, social context, personality, emotions, fatigue, and much more, affect the intent and the performance. In this context we can state that each motion can be performed in a high variety of ways, and it is challenging to exactly reproduce the same movement twice [Peng et al., 2014].

Luckily, our brain is capable of clustering and interpreting the intrinsic complexity of each perceived human motion, breaking down intentions and emotions [Mar, 2011]. For these reasons human movement is one of the major interpersonal **communication** medium but also one of the most delicate. Indeed, any form of communication is based on an agreed convention between two parts, defined based on a set of rules and a dictionary of messages, which we are more or less aware of. For certain kinds of communication means, like written languages, the rules are stricter, but the interpretation still depends on various factors, like the context or the cultural background. Though, the study of the characteristics of human motion is relevant in various fields, from medical and behavioral studies to performing arts, such as dancing or acting. This form of unspoken communication is included in the definition of “non-verbal communication”.

Non-verbal communication

"Non-verbal behavior refers to actions as distinct from speech. It thus includes facial expressions, hand and arm gestures, postures, positions, and various movements of the body or the legs and feet." [Mehrabian, 1971]

While facing another human we exchange a wide set of non-verbal information. Indeed, the communication message is the synchronous combination of multiple elements, where everything might be relevant, and anything that deviates from social conventions might alert us.

Communicating with virtual humans

Therefore, when we face a virtual human we expect to interact in the same way as we interact with our real counterparts [Hoffmann et al., 2009]. This expectation is higher if the human looks realistic, which means that, in this case, we tolerate less motions imperfections (this perceptual phenomena is called the "uncanny valley" and was defined by Masahiro Mori in [1970] for the study of humanoid robots). Since the generation of the first virtual human, proposed by the team led by William Alan Fetter in the early 1960's at Boeing industries in Seattle [Fetter, 1982; Hickey et al., 1968] to optimize reachability of aircraft's controls in an airplane's cockpit (see Figure 4), the gap with the real counterpart has been reduced both in terms of visual appearance and movement generation.

Indeed, the employment of virtual humans in interactive applications, requires a wide set of technical skills. Primarily, as we said, the synthesis of realistic appearances and natural motions, often the cause of the uncanny valley if they are not properly executed. But also, the development of complex reactive systems to adapt these motions to the user, *e.g.* in video games where the user controls the virtual character (which has to move accordingly to the provided inputs) and interacts with other autonomous characters (which are often required to behave and react realistically to events). Finally, the last requirement is to provide systems and means of interaction, and fruition of the virtual environment where virtual humans interact.

We can identify two main evolutionary steps regarding forms of interaction with virtual humans: (i) The first forms of interactions were unidirectional with virtual humans appearing on a visual display, *e.g.* the computer monitor for simulation, in movies and

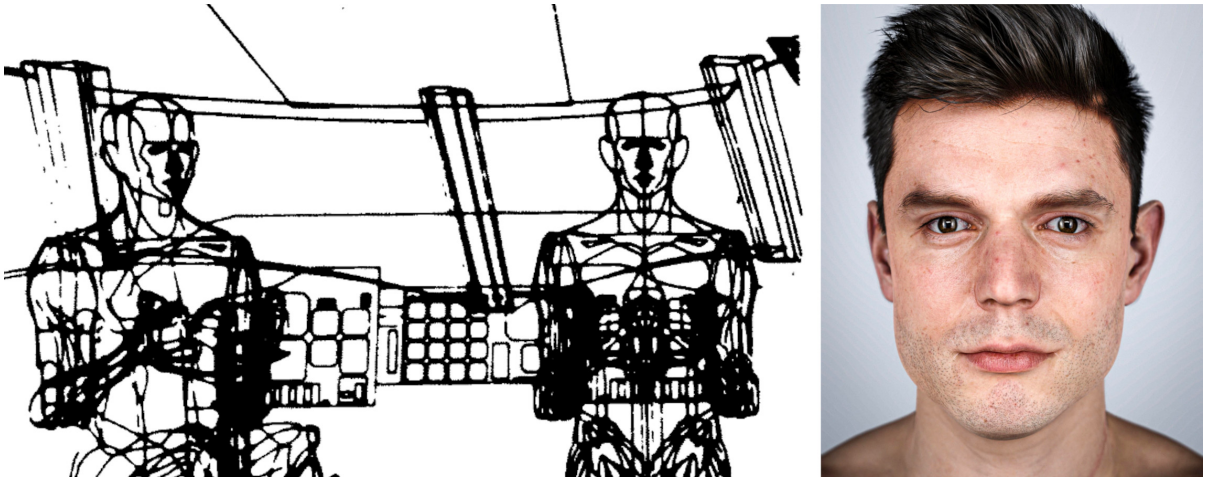


Figure 4 – Evolution of virtual character apperance: on the left, the "'First Man" [1968], a seven segment articulated pilot used in the development of Boieng 747; on the right, the character developed for the PRESENT project by Framestore [11] and Cubic Motion [12].

television advertisements. (ii) After that, with the evolution of interactive technologies, virtual humans required reacting capabilities, like for video games and immersive interactions like **virtual reality.**

Virtual reality

"Virtual Reality is a technical and scientific area making use of computer science and behavioral interfaces in order to simulate the behavior of 3D entities in a virtual world that interact in real-time among themselves and with the user in pseudo-natural immersion through sensory-motor channels." [Arnaldi et al., 2003]

In fact, virtual reality must actively engage one's senses to make the human user feel truly part of the virtual world. In this context, we evaluate the *immersion* in a virtual environment, by the distinct types of sensory information that flow from the virtual environment to the human user, and their similarity to real ones. Various technologies are involved in the simulation of these sensory feelings, notably head-mounted displays, for visual feedback, and haptic devices, for tactile simulation.

11. https://www.framestore.com/
12. https://cubicmotion.com/

Virtual Reality is often associated with an active user, interacting directly in the virtual world from his/her own point of view. However, when the user is not directly involved in the events, a different form of experience is required to ensure that he/she is appropriately presented with events. In this context, the main reference is cinematography. Indeed, from the conception of classical cinematography, a specific form of communication was developed, *i.e.* a visual grammar with rules for framing scenes, transiting between sequences – called cuts – and expression of visual styles, to convey narration and emotions to an audience [Bowen, 2013; Thompson and Bowen, 2009]. As audience members, we assimilate and share these conventions, which lead us to the capability of interpreting the intentions of the narrator. Actually, we are no more afraid of a train that moves towards us in a movie theatre, neither do we try to answer to the greetings of a journalist on television. In these terms, we can say that we developed new capabilities to communicate with shared communication schemas. Framing and compositions in movies are design in relation with the movement of performers, and the performance is adapted to the movement of the camera. When these rules are adapted for the fruition of virtual environments we talk about **virtual cinematography**. Virtual cinematography studies and proposes techniques to frame and illuminate environments to convey cinematographic experiences of events, such as interactions of virtual humans.

Thesis overview and Contributions

This thesis explores how to efficiently convey different non-verbal characteristics of human motion and focuses the attention on the various aspects of communication exchange between real and virtual humans. Indeed, each human motion conveys information, and if we want to simulate and reproduce more engaging and faithful motions for virtual experiences, we cannot prescind from the study of the communication forms and the perceptual aspects of motion. As mentioned, there are various forms of unspoken information exchange between humans, which, in virtual applications, are conveyed through different media from the virtual world to the human recipient of the communication. Through this manuscript, we adopt various terms to define this recipient, which can be a passive observer, an active interactant, a spectator, an audience, or a participant immersed in the virtual world. The choice of definition depends on the corresponding form of communication presented in the relative contribution.

This work presents a transversal exploration from different disciplines (human motion simulation, applied perceptions, multimodal immersive experiences, virtual cinematography, *etc.*) with the goal of enhancing the simulation of virtual humans in entertaining and interactive applications. Our work research was developed over four research axes.

Research axis 1: *to propose efficient animation methodologies in order to convey non-verbal characteristics of virtual human motion to the users.* For this case, our first contribution is an on-line edition of the upper body motion of a virtual character to fit user-defined visual targets from the point of view of an observer or virtual reality user. This means, for example, being able to adapt the orientation and amplitude of a waving motion of a character in relation with the apparent amplitude that the observer perceives from his/her position. If the observer is far, the agent needs to wave more to increase its visibility. To achieve this goal, we introduced a new paradigm that puts the observer of the motion at the center, towards whom the animation is performed, and adapts the virtual character's motion to satisfy various constraints expressed from the point of view of the observer.

Research axis 2: *to understand which factors affect the communication and how they are perceived by the users.* On this context, we investigate the perception of non-verbal clues of communication. In particular, we present two studies: (i) the first study evaluates the effect of upper body motion variations in the understanding of the virtual character intentions, in relation with the first research axis. (ii) The second one explores the effect of directed and averted gazing of a virtual human towards a human participant, a phenomenon called the stare-in-the-crowd effect which was observed in real life but not explored in virtual reality.

Research axis 3: *to study technologies that simulate the flow of sensory information from the virtual world to the user.* Our focus, on this case, is on the simulation of touch. We perform a user study evaluating how the sensory feedback simulating contact can affect the way we navigate through a virtual crowd.

Research axis 4: *to propose a technical solution in order to convey real-time events, involving virtual humans, as a cinematographic experience to an audience.* In this context, we examine how to portray previously unknown human motions in a cinematographic way,

by the analysis of the environment where they interact. We have designed a technique for placing cinematographic cameras and cameras trail through a topological analysis of the virtual environment. The goal of this work is to optimize the generation of cinematographic shots to follow the movement of characters and events in a virtual environment.

Thesis outline. Following these research axes, we describe here the structure of this manuscript.

In Chapter 1 we introduce the main concepts and review the literature related to the presented works. This first chapter is structured as follows. In particular, we discuss the main concepts related to the study of perception in interactive populated virtual environments, with a focus on non-verbal characteristics of the motion and how to convey these characteristics in a cinematographic way.

The following chapters present the contributions of this work.

In Chapter 2 we detail the first contribution of this manuscript, related to the first and second research axis. We demonstrate this technique over a set of use cases and a user experience in virtual reality (second research axis), to test if our technique can help the user to better understand the intention of the virtual agent.

In Chapter 3 we present the second contribution of this manuscript, related to the second research axis performed in virtual reality. In this contribution, we evaluate the presence of a perceptual effect induced by directed gaze in comparison with averted gaze while observing a virtual crowd. We introduce the technical set-up, the user-study and a discussion of the results.

Chapter 4 details the third contribution, related to the forth research axis, in which the evaluation focuses on the user capabilities of navigating through a virtual crowd. We study whenever the simulation of contacts through vibrotactile haptic feedback helps users performing such a task.

In the final contribution (Chapter 5) we move the focus from an internal interactant to an external observer of the interaction (forth research axis). In this chapter, we present a novel approach for positioning cinematographic cameras in virtual environments in order to follow the events, actions, and interactions of virtual characters in real time.

The Chapter Conclusion concludes this manuscript, by presenting the final remarks, a general discussion of the work presented over the thesis, as well as comment the overall future directions.

Publications related to this thesis

- **Alberto Jovane**, Pierre Raimbaud, Katja Zibrek, Claudio Pacchierotti, Marc Christie, Ludovic Hoyet, Anne-Hélène Olivier, and Julien Pettré. "Warping character animations using visual motion features." Computers & Graphics (2022). DOI: 10.1145/3424636.3426892
- Pierre Raimbaud, **Alberto Jovane**, Katja Zibrek, Claudio Pacchierotti, Marc Christie, Ludovic Hoyet, Julien Pettré, and Anne-Hélène Olivier. "Reactive virtual agents: A viewpoint-driven approach for bodily nonverbal communication." In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 164-166. 2021. DOI: 10.1145/3472306.3478351
- Pierre Raimbaud, **Alberto Jovane**, Katja Zibrek, Claudio Pacchierotti, Marc Christie, Ludovic Hoyet, Julien Pettre, and Anne-Hélène Olivier. "The Stare-in-the-Crowd Effect in Virtual Reality." In 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 281-290. IEEE, 2022. DOI: 10.1109/VR51125.2022.00047
- Florian Berton, Fabien Grzeskowiak, Alexandre Bonneau, **Alberto Jovane**, Marco Aggravi, Ludovic Hoyet, Anne-Helene Olivier, Claudio Pacchierotti, and Julien Pettre. "Crowd navigation in vr: exploring haptic rendering of collisions." IEEE Transactions on Visualization and Computer Graphics (2020). DOI: 10.1109/TVCG.2020.3041341
- **Alberto Jovane**, Amaury Louarn, and Marc Christie. "Topology-aware camera control for real-time applications." In Motion, Interaction and Games, pp. 1-10. 2020. DOI: 10.1145/3424636.3426892

Chapter 1

Background

Contents

In this chapter, we present a general overview of the generation, analysis and visualization of human motion. Section 1.1 explores the synthesis of motion for virtual humans. Then, in Section 1.2, we detail various studies on the perception of motion: *e.g.* gesture, gaze, touch and displacement. In this context, we explore how immersive technologies can help simulate these multimodal stimuli. In particular, we focus on the non-verbal characteristics of human interaction (Section Context of the Chapter Introduction). Finally, Section 1.3 introduces the concept of virtual cinematography, and so the control of camera to portrait the actions of digital humans.

1.1 Human motion synthesis

We define as human motion synthesis any process dedicated to the creation of digital movement of animated virtual humans [Guo et al., 2015; Van Welbergen et al., 2010].

Before diving into the generation of movement, it is important to remember how virtual human are commonly represented. A 3D mesh expresses the body appearance of virtual characters. Where, the triangles of this mesh are commonly linked – through the skinning process – to a high-level structure that controls the movement of the virtual body. One of the most used structures is a skeleton representation, *i.e.* a hierarchical composition of rigid bodies – called bones – connected by joints. These joints are stored as a rotational relation to their parent in the hierarchy and are parametrized either as Euler angle, rotation matrices, or unit quaternions.

This representation enables us to parametrize the virtual human kinematics as a sequence of frames, each displaying the current skeleton's configuration. The frequency of stored frames – key-frames – is balanced to optimize the trade-off between the level of detail and memory consumption. The additional in-betweens are then generated with the interpola-

tion of consequent key-frames, generating a continuous movement. Multiple approaches are proposed to synthesize human motion. In the following of this section, we will present the 3 main family of approaches, namely the generation of examples, the simulation of human motion through mechanics, and finally, the synthesis of new motions from existing examples.

1.1.1 Generating examples

We define as an example any stored animation data. In this section, we expose the main techniques used to generate motion examples.

Hand-made. Originally, motions were designed entirely by animators, who had to draw the temporal-spatial evolution of the skeletal structure at each keyframe. This approach, still in use for single-scope animation like animated movies, is not suitable for the workloads of the interactive applications, which require many animation variations. Because of this, research has evolved with tools to synthesize whole-body movements and to reduce artists' workloads.

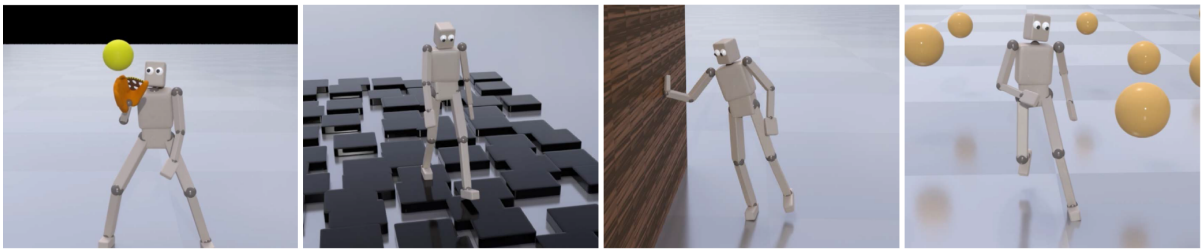
Motion capture. Motion capture is one of the most used techniques to synthesize animations. We define motion capture (mocap for short) as any process of digitizing an actor's movements using sensors. Since its inception, various technical solutions have been proposed. Depending on the type of sensors, we can distinguish two main categories: marker-based (acoustical, mechanical, magnetic, optical), more accurate but also more intrusive, and marker-less [Sharma et al., 2019], which include all the vision based methods [Colyer et al., 2018; Moeslund et al., 2006]. Motion capture is essential for live recording in immersive applications (*e.g.* embodiment [Genay et al., 2021]) and motion analysis (*e.g.* to evaluate athletes' performance [Van der Kruk and Reijne, 2018], or for medical rehabilitation [Zhou and Hu, 2008]). It is also valuable for populating motion databases, as mocap produces realistic and complex motion with short processing time. The principal drawback is that setting up a motion capture pipeline can be expensive and requires performing actors. Also, depending on the system, the data produced can be noisy and demand additional post-processing or artist's refinements.

1.1.2 Simulating human motion

The main alternative to recording or drawing motion examples is to use models to simulate them.

Kinematics. The main idea of this class of techniques is to reduce the required dimensionality of the manually-set degrees of freedom. With the definition of a few kinematic constraints (*e.g.* foot contacts, hand and head positions), we can infer the global configurations for the character's skeletal structure and, in extended cases, the transitional motion. These techniques belong to the *inverse kinematics* branch. *Inverse kinematics* defines the mathematical approaches, analytic [Unzueta et al., 2008] and numerical [Aristidou and Lasenby, 2011; Aristidou et al., 2016; Buss, 2004], applied to solve the under-constrained problem of extracting numerous unknown degrees of freedom from a set of defined ones [Aristidou et al., 2018].
Synthesis of full-body animations, through kinematic constraints, is complex because of the large size of the solution space, and additional constraints are therefore required to prune the results (*e.g.* energy minimization). For this reason, *inverse kinematics* is mostly used as support for the example creation process, *i.e.* for a low-dimensional mocap system, and as an artistic tool [Ciccone et al., 2019; Guay et al., 2015; Rose III et al., 2001]. Some of the analytic approaches and most of the numerical ones perform at interactive frame rates. On the negative side, the quality of the final result depends on the method and the provided parameters, which are not always easy to control.

Dynamics. Since early stages of animation different physical models were proposed to produce close-to-realistic movement. These approaches differ from kinematics ones because they simplify the biomechanics of the human body to simulate the connected masses of the musculoskeletal system, their control and the related forces [Shao and Ng-Thow-Hing, 2003]. The promise of physics-based character animation is to simulate a model that can act and react to external stimuli. Indeed, physics-based simulation easily succeed in this last task, *e.g.* to portrait inanimate character – *ragdoll phyisc* – while on the other hand still struggles to balance the generation of believable actions with real-time performances. The main advantage of physics-based approaches is to be able to interpret the interactions with the environment, such as external perturbations, and adapt to them.
Some examples show how various motions can be generated from the definition of



(a) Ball catching (b) Walking on stepping stones (c) Balancing by leaning on wall with one hand (d) Moving obstacle avoidance

Figure 1.1 – Examples of various gestures simulated with physics-based simulations from Eom et al. [2019].

a few constraints [Mordatch et al., 2012], or locomotion animations by setting the speed [Geijtenbeek et al., 2013]. These approaches are based on error-cost minimization, together with inverse dynamics (extracting controls and forces given positions and derivatives), and so they require a task-specific optimization time that might be heavy for variation in online applications. With some approximation and trade-off in quality, other approaches target online optimization for task-specific motions [Eom et al., 2019; Hämäläinen et al., 2014; Naderi et al., 2017].

These two simulation approaches deal with full-body motion generation with parallel philosophies with specific advantages and disadvantages. Kinematics-based techniques are typically more lightweight than dynamics-based ones, but, at the same time, they require more manually-set constraints, and they cannot autonomously handle external perturbations like dynamics ones. Both approaches fail in proposing an application-comprehensive model, but they often serve as support for hybrid techniques. *Inverse kinematics* is integrated in the majority of editing and tracking platforms, while dynamics is more often integrated in posture and motion adjustment, *e.g.* to simulate stiffness of links or balance of the center of mass, and interaction enhancements.

1.1.3 Example-based synthesis

With the growth of synthetic motion databases rose the need to modify and reuse animation data[Bodenheimer et al., 1997]. Motions can be adapted to new situations or can integrate artistic adjustments, and can also be combined to reduce the size of online animation databases. By definition, example-based synthesis uses previously generated example

motions to synthesize new ones [Wang et al., 2014]. We distinguish three sub-categories depending on the size of the input motions and the applied techniques: motion editing, interpolation, statistics-based. In the next paragraphs, we detail these sub-categories.

Motion editing. We define as motion editing the family of techniques that get one animation data as input and produce a variation of it. Editing approaches can affect the entire character motion or only sub portions of it. Early examples of motion editing expose control to kinematic constraints [Witkin and Popovic, 1995], apply signal analysis to filter motion curves [Bruderlin and Williams, 1995] and to represent emotional states [Unuma et al., 1995]. Follow-up researches extend the editing concept by performing dynamic time warping [Ashraf and Wong, 2000], others by shaping it on specific contexts, *e.g.* the environment [Gleicher, 2001; Ho et al., 2010],or contact interactions [Al-Asqhar et al., 2013], interpretation of emotion and personality [Chi et al., 2000; Durupinar et al., 2016]. Another form of editing, called retargeting, adapts animations to characters with various morphologies [Choi and Ko, 2000; Gleicher, 1998; Villegas et al., 2018]. In conclusion, the concept of editing is associated with tools for artistic refinements of animations data [Choi et al., 2016], but motion editing serves as efficient techniques to operate adjustments in any animation system.

Interpolation. As with the creation of the in-betweeners during the interpolation of two keyframes, new movements can be synthesized by coherently mixing together – blending – animations. Blending techniques not only deal with the mathematical interpolation problem (Bezier, B-spline, Euler-based, quaternion linear blending, spherical linear/spline quaternion interpolation, *etc.*), but also try to automate the generation of smooth believable motions out of a database. To do so researchers need to overcome various challenges: (i) the organization of the motion database to optimize the access [Keogh et al., 2004], as well as the synchronization and connection of sequences to avoid jitter [Adistambha et al., 2008; Kovar and Gleicher, 2004; Switonski et al., 2019; Wiley and Hahn, 1997; Zhou and De la Torre, 2012], (ii) the definition of high level controllers over the blending result [Kim et al., 2009], (iii) the definition of constraints and weights depending on the performed task and desired results [Ménardais et al., 2004]. Following these ideas, databases are reorganized into data structures. One of the first successful approach is the motion graph [Arikan and Forsyth, 2002; Casas et al., 2012; Heck and Gleicher, 2007; Kovar et al., 2002; Safonova and Hodgins, 2007] also optimized for tasks (*e.g.* interac-

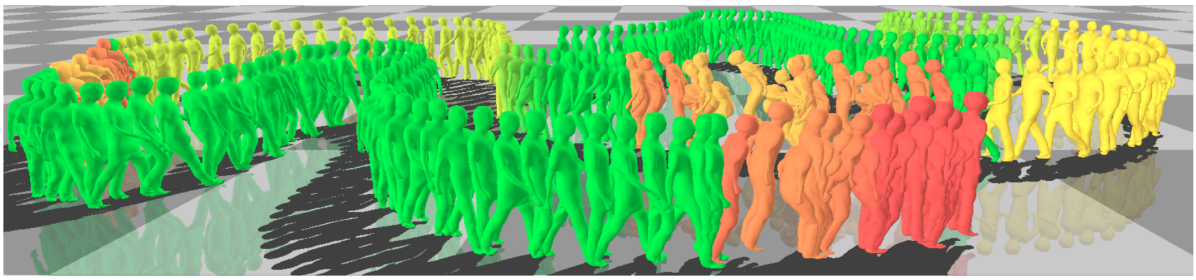


Figure 1.2 – Example of a sequence of motions interpolated in real-time from a database out of an high-level parametrization (highlighted by color variations), from Casas et al [2012].

tions [Shum et al., 2008]), followed by motion fields and motion matching [Arikan, 2002; Holden et al., 2020; Lee et al., 2002; Lee et al., 2010]. Others interpolate examples of the same kind to control kinematic features [Kovar and Gleicher, 2003; Park et al., 2004; Rose et al., 1998]. Interpolation techniques succeed in creating run-time accurate and fluid motions that are responsive and respect specific constraints or interpret high-level parameters [Randhavane et al., 2019a]. For this reason they suit well reactive applications like video games. The main drawback is that the efficiency and performances are directly proportional to the size of the motion database.

Statistics-based. Statistics-based approaches use a large motion database to define a statistical model for the synthesis of human motion. In recent years, techniques related to *machine learning* have grown in popularity. With this term, we define prediction tasks where the model parameters are learned from the data examples (*training phase*) and validated with new and unknown data (*testing phase*). There are two main categories of techniques: (i) *neural networks*, more commonly *deep learning* [Mourot et al., 2021] and (ii) *reinforcement learning* [Kwiatkowski et al., 2022]. In general, learning approaches tend to focus on specific domains of motions (*e.g.* locomotion [Feng et al., 2012; Glardon et al., 2004; Holden et al., 2016a; 2017b], interactions [Men et al., 2022; Starke et al., 2020; 2021]). An interesting branch of learning techniques explores the extraction and generation of stylistic variations. The base idea is that any gesture can be decomposed into two levels: the essential movement and its style [Unuma et al., 1995]. Therefore, researchers try to learn these style parameters from homogenous motion databases, *e.g.* style machines [Brand and Hertzmann, 2000], a statistical unsupervised learning technique, PCA-based (principal component analysis) [Urtasun et al., 2004] and style components [Shapiro

Figure 1.3 – Example of a motion and motion interactions generated with a statistics-based model, from Starke et al. [2021].

et al., 2006]. Style parameters can also be used to encompass aspects of personality or emotional state [Holden et al., 2017a], and be transferred from another motion [Liu et al., 2005; Xia et al., 2015; Yumer and Mitra, 2016] or from a video [Aberman et al., 2020].

A trained network can automatically variate and compose highly realistic animations online. Additionally, learning techniques can perform stylistic editing of motions, also called style transfer. The main drawbacks are that (i) the quality of the animation depends on the size of the motion database, (ii) the domain of the input motions defines the domain of the synthesized output – we can not generate a swimming motion from a database of running motions, (iii) learned parameters are often hard to interpret.

1.2 Interactive populated environments

Our interest is to reproduce realistic social behaviors in simulated scenarios, populated with autonomous humans. As humans, we voluntarily and involuntary participate in a continuous exchange of social interactions in everyday life. Therefore, it is crucial to translate these social characteristics to virtual humans. In the previous section, we explored how motion can be synthesized and adapted to various situations. In this section we detail the main challenges of simulating human social interactions in virtual environments. One part of these challenges is related to the simulation of the virtual world: how technologies evolve to generate populated synthetic environments. Another part is related to the definition of perceptual characteristics: as we mention in the introduction, our focus is the comprehension and the quantification of non-verbal cues, how these are interpreted by humans and how their perception varies from the real to the virtual context [Bailenson

et al., 2003; 2005; Bühler and Lamontagne, 2018; Narang et al., 2016].

1.2.1 Perceptual metrics

The technical evolution of virtual reality simulations facilitates the digital generation of faithful real life situations. Indeed, the evaluation of a virtual environment starts with what we are able to simulate. The level of immersion can be objectively measured [Slater, 2003], and depends on the technology stimulative power to replicate real world senses (*e.g.* headset, haptics device). In addition, subjective measures can be accounted to evaluate the perceived "realism" of immersive environments. These measures are compared on similar immersive conditions. *Presence* was introduced by Slater and Usoh [2003; 1993] to describe the sense of being in the virtual place, by comparison with the real one. It is now widely demonstrated that the level of *presence* felt by users is affected by the interaction capabilities of the virtual humans [Bente et al., 2008; Schuetzler et al., 2018; Von der Pütten et al., 2010], this branch of *presence* study can be defined as *social presence* [Oh et al., 2018]. An additional metric is related to the measure of how we perceived the non-physical body as our own, that is called *virtual embodiment* [Kilteni et al., 2012]. Furthermore, authors have shown the ecological validity of virtual reality to reproduce real-life non-verbal communication human behaviors [Li et al., 2019; Pan and Hamilton, 2018]. In this context other relevant social metrics can be evaluated, such as *social agency* [Silver et al., 2020], the sense of how much one's own actions can affect the populated environment. Another relevant factor is the study of social discomfort in virtual reality, such as *social anxiety* [Clark, 1995] which defines one's fear of negative evaluation from the side of others. Such anxiety can be reflected by body cues such as heart rate increase, both in real-life situations [Pittig et al., 2013] and virtual ones [Kahlon et al., 2019], and it can be caused by non-verbal cues.

To conclude, perceptual metrics are relevant to evaluate the naturalness of users' interactions with virtual humans [Allmendinger, 2010; Hodge et al., 2008] and to generate compelling and realistic sentient virtual humans.

1.2.2 Non-verbal characteristics

As we discussed in the Introduction, the focus of this manuscript is to study communicative interpersonal exchanges not related to spoken language [Burgoon and Bacue, 2003]. These non-verbal exchanges rely on various interaction aspects, notably proxemics

and kinesics, where posture and motion of the head, body, and limbs, touch and gaze behaviors are all relevant cues of investigation [Harrigan et al., 2008]. In the following, we will present works from four of these non-verbal aspects.

Kinesics. Kinesics refers to the study and interpretation of human body motion. The body is one of the main medium for non-verbal communication. Through the motion of limbs and the posture of a person we can already recognize cues of its emotional state, its attitude and its personality. Identifying quantifiable patterns in the body motion is challenging, and the interpretation of the behavior is not always objective or unique. Various schemes were proposed for low-level motion characteristics, the most famous being the classification introduced by Laban and Ullmann [1971], that leads to the decomposition of motion in four categories: body, effort, shape and space. The Laban notation system serves as a basis for various classification methods [Bouchard and Badler, 2007; Durupinar, 2021] and motion synthesis tools [Chi et al., 2000; Durupinar et al., 2016; Garcia et al., 2019]. Other behavioral protocols are the Bernese system [Bente et al., 2001; Frey and Von Cranach, 1971] coding variations from "normal" positions, and the Birdwhistell system [Birdwhistell, 2010] that mimics linguistic principles. Dael et al. [2012] approached the classification of gestures, inspired by the *facial action coding system* [Ekman and Friesen, 1978; Hager et al., 2002] a well-known concept for facial animation, defining body action and posture units as instances of body motions. Furthermore, the visual perception, in social interaction, enables to infer on a person's intention [Blakemore and Decety, 2001; Knoblich and Sebanz, 2008; Perrinet et al., 2013], personality [Neff et al., 2010; Smith and Neff, 2017], as well as emotions [Ahmed et al., 2019; de Gelder et al., 2015; McDonnell et al., 2008; Randhavane et al., 2019b; Roether et al., 2009]. In conclusion gestures and body postures are varied and complex, but vital for the believability of human interactions. From a perception point of view, the construction of flexible models to simulate interactive behaviors is one of the major challenge the character animation community will face in the next years.

Proxemics. Proxemics defines the study of the variations in interpersonal space between individuals. Indeed, another important aspect in social interactions is how we move in relation to others. Proxemics studies characteristics such as how far we position ourselves and move from another person, or how the perception of our personal space mutates. Iachini et al. [2016a] showed that proximity behaviors to virtual agents in virtual reality

resembles the characteristics people exhibit in the real-world. Proxemics behaviors are influenced by cultural [Hall, 1969] and demographic aspects [Iachini et al., 2016b; Zibrek et al., 2020], as well as by the setting [Duverné et al., 2020] and by other non-verbal behaviors [Bönsch et al., 2018]. These factors are amplified even more when we consider crowds, and so understanding how multiple humans interact became a fundamental requirement for the design of realistic situations. To that purpose, several studies have been conducted to investigate collective behaviors. For example, Seyfried et al. [2005] showed that speed depends on density in a crowd (i.e., fundamental diagram) and this relation is affected by cultural aspects [Chattaraj et al., 2009]. Bonneaud et al. [2012] showed that, without any instruction, collective behaviors can emerge within a group of walkers and these behaviors can be described according to several patterns, such as the anisotropy of interpersonal distance or speed synchronization. Some other studies focused on the local aspects of interactions. Using pairwise situations, Olivier et al. [2012] previously showed that pedestrians adapt their motion only if there is a future risk of collision. Others considered the effect of situational factors such as crossing angle [Huber et al., 2014; Knorr et al., 2016], crossing order [Olivier et al., 2013] or orientation [Bourgaize et al., 2020] as well as personal factors such as gender and personality [Knorr et al., 2016] or body size [Bourgaize et al., 2020] on motion adaptations. While these previous studies have considered the kinematics of the adaptions, other works were interested in the gaze activity, showing that it can predict future crossing order [Croft and Panchuk, 2018] and that gaze behaviors are task-dependent [Hessels et al., 2020].

Researchers agree that virtual reality is very promising for conducting such experiments, since the nature of the interactions is preserved and participants are expected to behave in the same way as in real conditions. A lot of effort has been put to validate this approach, both for the study of trajectory [Olivier et al., 2017; Silva et al., 2018] and gaze [Berton et al., 2019] behaviors. However, some quantitative differences have been observed, *e.g.*, an increase of the crossing distance and larger head movements, which can be due to the distorted perception of distances and limitations of the field of view introduced by virtual reality head-mounted displays. Nevertheless, virtual reality opens large perspectives in the design of new experiments for understanding pedestrian behaviors. For examples, having only one participant at the same time, researchers were able to evaluate the effect of specific factors such as crowd emotions on proxemics [Bönsch et al., 2018; Huang and Wong, 2018; Volonte et al., 2020]. While the interest of virtual reality is widely established for the study of proximity interactions, studies in virtual reality have

mainly designed experimental paradigms which involved vision, proprioception as well as the vestibular system to perceive user actions and their surroundings in the virtual environment.

Touch. Touch is a primal sensation to analyze our surroundings. While the exact reproduction of tactile real-life stimuli in virtual environments is currently unfeasible, contact situations can nevertheless be simulated with a wide range of haptic devices. When we think about touch we typically associate it with the hands. *E.g.*, Pacchierotti et al. [2017] presented a review paper on wearable haptic devices used to render contact sensations at the fingertip and hand. Notable examples employed this technology in virtual reality applications, both to enhance the perception of contacts and stiffness of materials [Chinello et al., 2017b; 2019; Salazar et al., 2020] and also as guide for navigating virtual environments [Schorr and Okamura, 2017]. But contacts are, of course, not limited to the hand. In this respect, Lindeman et al. [2004; 2006] developed a wearable haptic vest and belt capable of providing distributed vibrotactile feedback sensations in virtual reality. Vibrations were used to indicate a contact in the virtual environment or to provide information about areas of the environment yet to explore. Indeed, the simulation of contact in virtual reality has been extensively studied for applications involving more than just contact. For instance, Mestre et al. [2016] used a vibrotactile haptic device for rendering proximity to obstacle during avoidance task in virtual reality. Louison et al. [2018] showed that wearable vibrotactile devices increase spatial awareness and reduce collisions in an industrial training scenario, while [Aggravi et al., 2018; Bimbo et al., 2017] employed vibrotactile devices to provide feedback while operating a robotic arm. Boucaud et al. [2021] use wearable haptics to design a system for mutual touch between a participant immerse in virtual reality and a virtual human, with the promise of increase the believability of a social interaction. Regarding interaction with virtual characters, Krogmeier et al. [2019] designed an experiment where participants had to bump into a virtual character, with or without haptic rendering of contacts. This haptic rendering was performed using the "Tactsuit" [1], equipped with 70 haptic points of contact. In this preliminary study, they showed that this kind of haptic feedback improves presence and embodiment. In another context, Krum et al. [2018] were interested in the impact of different locomotion techniques and priming haptic rendering on proxemics and subjective social measures during interactions with a virtual character. The priming haptic rendering corresponded to a sim-

1. https://www.bhaptics.com/

ulated touch by the virtual human. Their results showed that priming haptic rendering did not influence participant's proxemics but influenced the subjective social measures. For instance, it improved the sympathy and the relation toward the virtual character. Furthermore, Faure [2019] asked participants to perform a collision avoidance task with a virtual character, while walking on a treadmill and with haptic feedback. Additionally, touch can be used to simulate social and emotion feeling, Teyssier et al. [2020] study the emotional response to various touch pattern of an artificial hand, attached on a robotic arm, on the participant forearm.

Eye's gaze. Gaze behavior plays a primal role in interpersonal interactions, and in the same way, when we relate with a virtual human or a humanoid robot. To better capture such gaze behaviors, Studies focus on two main tasks: (i) studying the perception of gaze and (ii) simulating synthetic gaze behaviors [Ruhland et al., 2015]. During social interactions, a continuous exchange of such signals is made possible mainly because people are able to see each other [Cañigueral and Hamilton, 2019]. For gaze behaviors in dyadic interactions, Bailenson et al. [2001] showed the preservation of the equilibrium between mutual gaze and personal space distance in virtual reality. Additionally, Garau et al. [2003] showed the effect of an inferred-gaze model on perceived quality of communication in virtual reality, compared to a random-gaze model. In line with this, Nummenmaa et al. [2009] showed the importance of virtual reality users' interpretation of virtual agents' gaze cues in order to avoid collisions when navigating towards them. For user-virtual agent interactions in the context of a crowd, Narang et al. [2016] also confirmed the importance of the modelling of gaze interactions, reflected by an increase of the believability of the interaction when comparing it with and without using gaze models in the virtual environment.

To conclude, previous studies have shown the importance of gaze communication during interactions between a user and virtual agents, where virtual reality is able to preserve real world behaviors, as assessed by the social behaviors of virtual reality users among a virtual audience [Iachini et al., 2016a; Nummenmaa et al., 2009] and by users self-assessments such as presence and engagement [Chollet and Scherer, 2017; Glémarec et al., 2021; Roth et al., 2018].

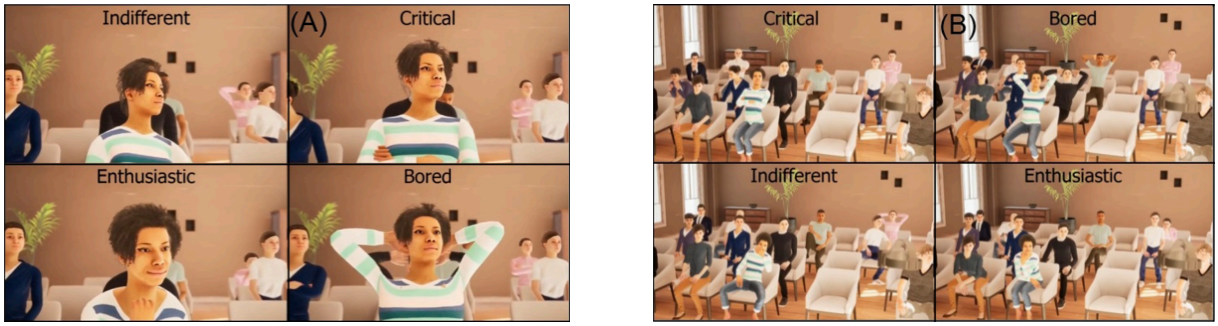


Figure 1.4 – Example of visual stimuli for a study of virtual audience perception in virtual reality, from Glémarec et al. [2021]. On the left (A) for a single agent, on the right (B) for the entire virtual audience.

1.3 Camera control in virtual cinematography

After we explored how non-verbal communication is generated and perceived, and how humans are able to interpret others' motions, we deal, in this section, with the extended problem of how to optimally display virtual human motions and interactions. We present this problem within the more general concept of camera control for virtual cinematography.

With virtual cinematography we define the application of the cinematographic language [Bowen, 2013; Thompson and Bowen, 2009] in virtual environments, *e.g.* for the control of illumination and cameras [Debevec, 2006]. Camera control deals with issues in placing and moving cameras in virtual environments [Christie et al., 2008], to frame motion, actions and interactions between characters. It is a well-studied problem in computer graphics, and approaches have been exploring how visual features such as target visibility, screen composition, optimal view or camera smoothness can be enforced [Christie and Olivier, 2006] by relying on motion-planning, optimization and more recently deep-learning techniques [Jiang et al., 2020; 2021]. In the context of real-time 3D applications such as game engines, contributions have essentially focused on target tracking techniques [Halper et al., 2001], coupling visibility with path-planning techniques [Oskam et al., 2009] or evaluating it in real-time [Burg et al., 2020].

1.3.1 Automated generation of camera paths

The computation of camera paths with a prior knowledge of the environment is either performed as a path planning process or a motion planning process (*i.e.* integrating temporal information). Different planning techniques have been proposed to guide the motion of cameras based on the underlying representations proposed in the robotics literature (see [Lino et al., 2010; Oskam et al., 2009]). For example, Oskam et al. [2009] relied on a prior spherical decomposition of the free-obstacle space, by filling the space with intersecting spheres. Visibility between each pair of spheres is also precomputed using ray-casting and stored. A graph-based roadmap of the environment can then be constructed where each node is the center of a sphere, and each edge is a collision-free motion from one sphere to the neighbor intersecting one. At run-time the roadmap is queried with an A* algorithm to compute the shortest path from the current camera position to the target position that maximises the visibility of a target. To highlight the motion of a vehicle, Huang [2016] rely on an interactive optimisation technique which computes a sequence of waypoints that will ensure the proper tracking of targets by the camera. Key characteristics to optimize are visibility of the target, camera smoothness, and visual load (the more objects in the scene, the slower the camera is).

The key issue common to most path or motion planning approaches is actually how to characterize what makes a good cinematographic motion. While smoothness (expressed as the absence of jerk on the evolution of camera trajectories) is often considered, there is no clear consensus on characteristics of good cinematographic motions. Galvane et al. [2015a] therefore proposed to create camera trajectories by performing interpolations of cinematographic properties in the screen space (*e.g.*, angle on targets, composition of targets on the screen, size of targets). Later, they exploited the idea further to create camera paths for drone motions that would avoid sudden on-screen changes [2018].

Most approaches however only focus on the computation of a single camera, or camera path, to perform the requested task, and have not been addressing the issue of populating environments with cinematographic cameras.

1.3.2 Maximum coverture issue

As far as we know there is little literature available on the automated placement of cinematographic cameras in 3D environments, driven by the topology of the environment. Some previous work address the issue of automated camera placement typically in the

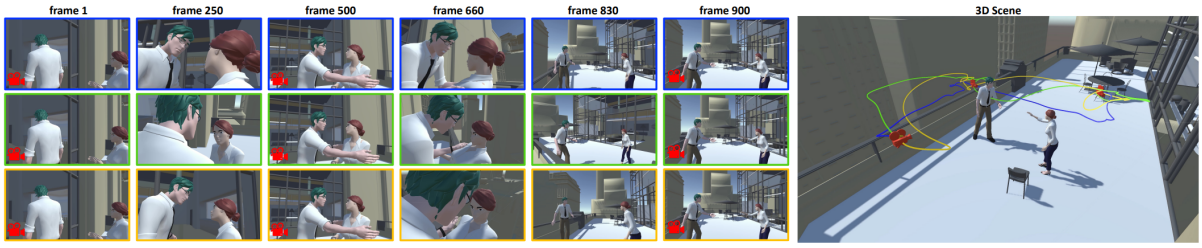


Figure 1.5 – Example of a various sequence of camera motion generated automatically by the technique proposed by Jiang et al. [2021]. Each sequence (blue, green and yellow) is generated with a different behavior while respecting the given constraint, represented by the red camera position and orientation.

context of the Art Gallery problem [O'Rourke, 1987]. This is a well-known optimization problem where the goal is to place the minimum number of surveillance cameras to cover the entire surface of an art gallery. Or instance, van den Hengel et al. [2009] solved this problem using a genetic algorithm to place the cameras given a 3d model. Other approaches, such as Chittaro et al. [2010], proposed the design of an authoring tool that generates virtual tours to ease the navigation process, yet the specification of POIs (points of interest) is defined manually. On our side, we are not interested in the minimum number of cameras, nor a limited set of POIs but to obtain qualitative views on some possible events inside the scene, in particular related to virtual characters actions, which are not known beforehand. The particular challenge we face here is the computation of camera locations and motions without knowing beforehand the motion of characters and events.

In conclusion, multiple processes have been proposed to generated camera paths, with the goal of framing the actions in a cinematographic way.

1.4 Conclusion

In this chapter, we briefly explored the main concepts and techniques used to generate, study and frame virtual human motions. In recent years, with the advent of immersive technologies – in particular virtual reality – the need for more compelling animated virtual characters requires the development of realistic and real-time techniques to adapt motions, simulate interactions and synthesize reactions. As we saw, more studies are

moving in this direction, but still, few consider the user as an active part of the animation pipeline. Indeed, the communication between embodied users and virtual agents gain huge importance for future technologies and the announced advent of a virtual shared environment. In this context, how we perceive virtual humans and their motions is still a non totally solved problem. We have seen how delicate is the perception. Indeed, we are in the process of understanding these communication patterns. This will lead us to more accurate ways of generating digital humans, and a better way of communicating with virtual worlds, in the same way we do in the real one.

Chapter 2

Vision-based motion editing

Contents

In this chapter we present our first contribution to the run-time editing of the kinematic properties of a character motion in relation to the estimation of its characteristics as seen from the point of view of an observer.

We start by introducing the context that motivates this contribution. Numerous interactive applications, *e.g.* video games or Virtual Reality immersive simulations, rely on motion capture techniques to animate human characters thanks to the excellent trade-off they provide between computational budget and animation quality [Bodenheimer et al., 1997]. In the exposition of the related works (Chapter 1) we described various approaches for synthesizing motion, and also highlighted how the centrality of the user is, sometimes, neglected by those systems. Furthermore, only a few approaches have considered the influence of the observer, and the resulting visual features it yields, as a mean to control and warp a character animation.

Yet, establishing relations between visual features and motion characteristics of the animation is an interesting solution in the frame of non-verbal communication situations [Hinde, 1972] where vision is a major perception channel. This means that the interpretation of social signals involving a mutual interaction between the observer and the other(s) person(s) should be expressed in the reference frame of the observer, namely his/her field of vision. Indeed, in such communication tasks, humans control their movements by fundamentally taking into account how it can be visually perceived by an observer. For instance, while waving at someone – a typical voluntary non-verbal communication gesture to attract attention – one makes sure his/her hand is visible to this person, *e.g.* adjusting body orientation, waving amplitude and speed to make the motion salient enough, as well as moving his/her face and eyes to enable gaze contact and ensure that attention is successfully attracted.

This example clearly highlights the links between the kinematics of a motion and the visual features perceived by an observer. Previous approaches demonstrate how motions can be synthesized from the analysis of general visual properties, *e.g.* we see examples in animation, where images of the environment are exploited to generate character movements [Cao et al., 2020]. More specifically, a similar vision-optimization approach is presented by Huang and Kallmann [2015], where character motion, posture and location, are synthesized in relation to an observer. Despite proposing a comprehensive framework, this work is limited to a set of predefined actions and only considers visibility constraints as a mean for character placement. We believe that a finer definition of the character-observer relation is needed, so that motions and posture adaptation have a direct link with vision-based features.

Contributions. Our main contribution is a new visually-driven motion editing framework that enables to manipulate the visual features of an existing motion through an observer's viewpoint. This notion of *visual motion features* is defined in relation with an observer's point of view and field of vision. It covers features such as visibility and centrality of limbs, coverage of limbs and of their motion, generated optical flow, motion amplitude in the view plane, etc.

We express the problem as a specific case of visual servoing [Chaumette and Hutchinson, 2006], where camera feedback can guide the kinematic properties of the motion of robotic limbs. Visual servoing is a technique to control the motion of a robot based on

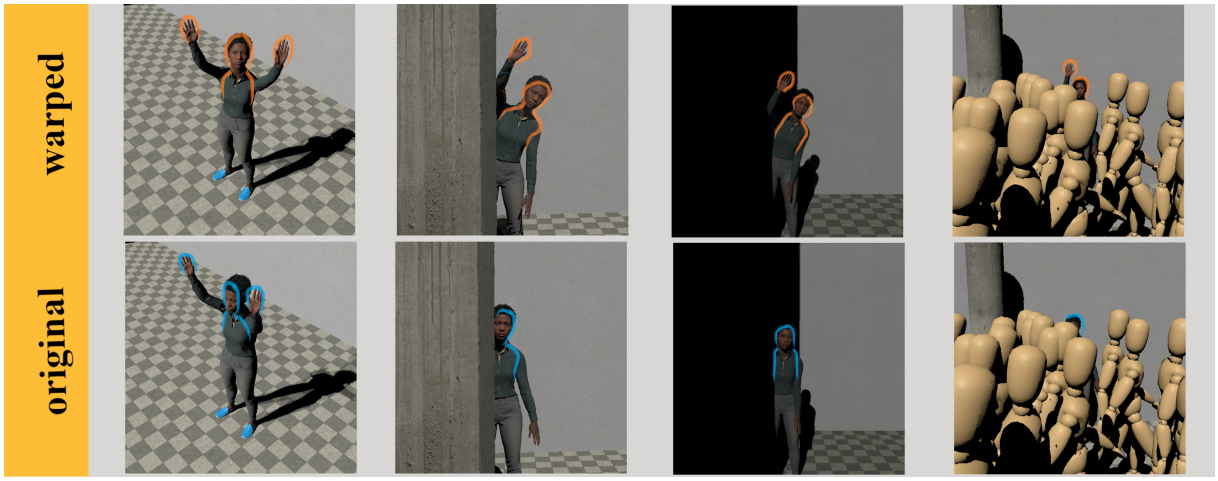


Figure 2.1 – We propose a novel and automated real-time motion editing technique that performs a view-dependent environment-aware warping of character animations driven by user-specified visual features. The bottom row images display examples of original animations and the top row displays the warped versions of the same animations. The warping process is driven by the user specification of visual features. We here illustrate how an increment in desired visual coverage impacts the kinematic chains of the character, and helps to draw more attention. Our visual features are also aware of visibility (second and fourth column) and lighting conditions (third column).

visual sensor feedback, *e.g.* a camera. In analogy with visual servoing, we want to control a character motion with respect to an observer position. Following this analogy, our motion warping goal is formulated as a visual task, *i.e.* to reach s^*, the visual target.

Unlike previous approaches where motion editing goals generally rely on the kinematic properties of the motion, we directly set the visual properties we want for the edited motion. However, a major difference with visual servoing solutions is that we want to perform motion warping, not full motion control. The objective of this work is to devise a motion editing technique that enables controlling visual motion features of a given character relatively to an observer. In other words, given a character motion m, the corresponding set of visual features s as perceived by an observer, and the *desired* visual motion feature values s^*, we search for the warping operation w to synthesize the warped motion m_w with the desired s^* values.

In Section 2.1 we detail how the technique works, then in Section 2.2 we present a set of use cases to illustrate how our technique can adapt existing content (motion capture) to a new context, and may empower virtual agents with the ability to capture attention. Additionally, we evaluate the relevance of the approach when animating characters in

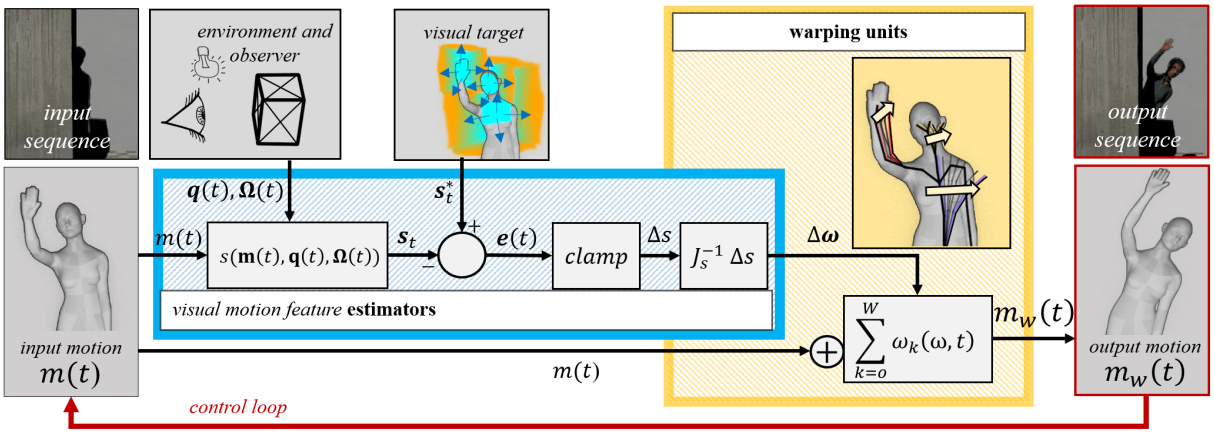


Figure 2.2 – Overview of our approach. From an input sequence of a character animation, we first estimate different visual motion features on the current pose, considering the environment, the observer's viewpoint, and a visual target (blue). Then, multiple plausible motion modifications are computed manipulating warping units (yellow), and the ones that minimize the visual error between the current state and the target are applied to the output motion. This process is repeated for the whole motion over a control loop (red).

social interactions, or in immersive applications where characters are meant to adjust themselves to the users' state in the scene. Thus, in Section 2.3 we present a user study in virtual reality where we demonstrate the advantages of our approach compared to a more standard technique. Finally, in Section 2.4 we discuss the limitations of our approach, as well as future works.

2.1 Method

In this section, we describe the mathematical formulation of our approach. Section 2.1.1 introduces the main concepts, Section 2.1.2 presents the visual motion features we consider when warping the animations, Section 2.1.3 details the design of our warping operators, and Section 2.1.4 explains the regulation of the visual motion features using our warping operators.

2.1.1 Overview

The aim of our approach is to provide designers or interactive applications with a warping controller that adjusts a character animation using view-dependent visual motion features. The input of our system is (i) an animated character, (ii) a set of designer-

specified visual motion features to fulfil (*e.g.* visual coverage, vertical extension), and (iii) a camera angle or trajectory that views the character in its environment. Our system adjusts the character animation on a per-frame basis by satisfying user-defined visual motion features in an inverse design approach, from features to parameters.

Technically, we propose to express the problem as a specific instance of the eye-to-hand visual servoing principle [Dombre and Khalil, 2013] in which a camera, fixed or animated in the world, observes the motion of one or multiple kinematic chains. Rather than driving the velocity of kinematic chains from on-screen velocities – as implemented in traditional visual servoing tasks [Espiau et al., 1992] or with through-the-lens control [Gleicher and Witkin, 1992], our objective is to develop a control law that updates velocities in the kinematic chain by regulating the difference between globally measured visual motion features and expected ones in the 2D camera space. This requires the design of (i) **estimators** that are able to measure the values of the expected visual motion features from the given camera, (ii) **warping units** that alter the parameters of the kinematic chains, and (iii) a **control loop** that exploits the difference between estimated and expected features to drive the warping operators. Our overall approach is described in Figure 2.2.

2.1.2 Estimators of visual motion features

The perceptual mechanisms by which humans look at, read, and understand images are well studied nowadays. Typical features such as chrominance, contrast and motion in the spectator field of view are well-known bottom-up key factors that influence audience attention. Attention is also driven by a number of top-down factors such as object semantics (faces draw strong attention), cultural background, and tasks to perform [Kimura et al., 2013].

With this in mind, we propose **estimators** to measure *visual motion features*, that are computational characteristics designed to measure how well the motion of a character is perceived from an observer's viewpoint. Moreover, these features represent here a proxy for visual attention. As such, they provide a view-dependent metric influenced by the character motion, the lighting in the scene and by potential occluders, and therefore provide a mean to control the amount of perceived motion in screen space.

For a given time frame t, we express the visual motion features as a time dependent vector $\boldsymbol{s}_t$:

$$\boldsymbol{s}_t = [\boldsymbol{s}_1(t), ..., \boldsymbol{s}_V(t)]^T = s(\boldsymbol{m}(t), \boldsymbol{q}(t), \boldsymbol{\Omega}(t)) \tag{2.1}$$

where $\boldsymbol{s}_i(t)$ is a function of *a)* the character pose specified by $\boldsymbol{m}(t)$, *b)* the observer pose specified as a camera pose $\boldsymbol{q}(t)$, and *c)* the state of the environment $\boldsymbol{\Omega}(t)$ that accounts for the rest of the scene, notably any lights and geometries that may affect the visibility of the virtual agent from the observer viewpoint. The function $s(\boldsymbol{m}(t), \boldsymbol{q}(t), \boldsymbol{\Omega}(t))$ performs the scene and character rendering from the given camera, and computes the visual motion features.

Visual motion features. In this work we consider the following features:

— *apparent static coverage* measures how much of a character's projected image is perceived in a frame (accounting for visibility and lighting);
— *apparent static extension* measures the horizontal occupancy (resp. vertical occupancy) as a ratio between the left-most and right-most pixels (resp. bottom-most and top-most) of the character on the screen width (resp. height), also accounting for visibility and lighting.
— *apparent motion coverage* measures how much of a character's motion from one frame to another is perceived in the image;

Likewise, designers are free to define and use additional visual motion features for their own application, at the condition that these can be represented as scalar values. In practice, visual motion features are computed through hardware rendering and straightforward image analysis. At each time step, a frame is rendered from the observer's viewpoint and only the *perceived* pixels of the virtual agent are kept. A pixel is considered as *perceived* if it is not occluded, or if its luminance is under a given threshold (*e.g.* in a shaded or dark area). Visual motion features are estimated through pixel operations such as counting or comparing coordinates and distances in the image space.

Semantic layers on body representations. Sub-meshes of virtual characters are tagged to identify specific parts (face, arms, chest, legs, inside of hands). This enables us to arbitrarily activate or deactivate body parts according to the performed motion, *e.g.* to focus and render only the waving hand in a waving character case. In addition, rigid objects can be attached to the skeletal joints and visual motion features can be computed on them (*e.g.* an additional piece of clothing or holding a sheet of paper).

2.1.3 Motion warping units

We first rely on a classical skeletal structure with joints using a tree of kinematic chains. An animation of the skeletal structure is defined as a set of keyframes along with an interpolation technique. We can thus define a function $\boldsymbol{m}(t)$ that computes the current pose of the character at time t, expressed as a vector of the degrees of freedoms of the character joints denoted $\boldsymbol{\theta}_t$:

$$\boldsymbol{\theta}_t = [\theta_0, .., \theta_K]^T = \boldsymbol{m}(t) \tag{2.2}$$

where K represents the number of degrees of freedom of the skeletal structure. Rather than regulating visual features by controlling simultaneously the whole vector $\boldsymbol{\theta}$ of joints of a character (which may create unexpected or unrealistic changes in the body poses), we propose to define specific groups of joints in the skeletal representation and define these as *motion warping units*. These groups are defined to provide a localised control on a character (*e.g.* only the spine, only the arms, only the head plus shoulders), which is a classical approach when designers need to locally warp motions without impacting the whole body (see Table 2.1). Furthermore, we design our motion warping units in a parameterised way that maintains the coupling between parameters of the kinematic chain by using a linear combination given a warping factor. This enables small warpings on the animations without loosing the nature of the motion.

Our motion warping unit is therefore defined as a function $\boldsymbol{\omega}_k(\omega, t)$ that, given a scalar displacement value ω, computes a sparse *pose offset vector* at time t affecting only the joint angles in the warping unit. For a given warping unit, the derived pose offset vector is added to the current pose $\boldsymbol{m}(t)$ to generate the warped animation. The index k represents the k-th warping unit, and $\boldsymbol{\omega}_k(\omega, t)$ yields a vector of size K with a zero-value for the degrees of freedom that are not offsetted. Our motion warping unit defines a pose offset vector in which each offset is a weighted linear combination of ω:

$$\Delta\boldsymbol{\theta} = \boldsymbol{\omega}_k(\omega, t) = [w_0^k\omega, .., w_K^k\omega]^T \tag{2.3}$$

The scalar value w_i^k is a weighting constant specific to a given warping unit k and degree of freedom i of the kinematic chain and can be viewed as a stiffness coefficient, traditionally used when manipulating inverse kinematic chains. These weights are manually set depending on the desired mobility of the related warping unit. The weights might be either statically defined in relation to the type of joints involved (for example, the neck joint has different mobility than the shoulder one) or adapted to the saliency of specific

motion, like the waving hand. The linear combination with ω ensures a coupling in the offset computation of animation parameters. The corresponding value $\Delta\boldsymbol{\theta}_i$ represents the i-th kinematic angle offset computed by the warping operator.

Each *motion warping unit* therefore computes a pose offset vector $\Delta\boldsymbol{\theta}_i$, and all offset vectors are aggregated on the current character pose to create the warped motion (see Equation 2.4). The magnitude and direction of the computed offset vectors are driven by a vector $[\omega_1, .., \omega_k]^T$ of warping unit parameters, the value of which is computed by the visual servoing task (see Section 2.1.4). The overall warped parameters $\boldsymbol{m_w}(t)$ of the animation at each time t are given by:

$$\boldsymbol{m_w}(t) = \boldsymbol{m}(t) + \sum_{k=0}^{W} \boldsymbol{\omega}_k(\omega, t) \tag{2.4}$$

2.1.4 Driving warping units through visual motion features

Our objective is to compute the optimal set of warping unit parameters $\boldsymbol{\omega}$ at each time t of the animation from a given set of desired visual motion features $\boldsymbol{s}_t$, through a **control loop**. We first express the relation $\boldsymbol{s}_t = f_t(\boldsymbol{\omega})$ where f_t computes the estimation of visual motion features. As a direct relationship between $\boldsymbol{s}_t$ and $\boldsymbol{\omega}$ exists, our goal is to solve this equation to obtain $\boldsymbol{\omega}$ from $\boldsymbol{s}_t$. Due to the strong non-linearity of the relation between visual motion features and kinematic parameters, a classical approach is to study the problem in the velocity space.

Given this set of visual motion features $\boldsymbol{s}_t$ that depend both on a camera viewpoint at time t and a set of warping unit parameters $\boldsymbol{\omega}$, the differential $\dot{\boldsymbol{s}}_t$ expresses how the variations in the visual features are related to the camera and the character animations.

Introduction on visual servoing. As defined, this problem is a specific case of a eye-to-hand visual servoing problem where a specified velocity in the image space of a fixed camera looking at a kinematic chain is used to drive its degrees of freedom [Espiau et al., 1992]. This visual servoing relation is generally defined as:

$$\dot{\boldsymbol{s}}_t = \boldsymbol{L_s} \boldsymbol{V}_n \boldsymbol{J}_n(\boldsymbol{\theta}) \dot{\boldsymbol{\theta}} + \frac{\delta s}{\delta t}$$

where $\boldsymbol{J}_n(\boldsymbol{\theta})$ is the Jacobian of the kinematic chain, $\boldsymbol{V}_n$ the kinematic tensor transformation from the camera to the character, $\boldsymbol{L_s}$ the interaction matrix, and $\frac{\delta s}{\delta t}$ describes the variations of $\boldsymbol{s}$ caused by a movement of the camera (n represents the number of degrees

of freedom of the kinematic chain). This relation defines the correlation between the variation in visual motion features and the variations in degrees of freedom of the kinematic chain. We rely on this formulation to express our problem in terms of the warping unit parameters $\dot{\omega}$.

$$\dot{\boldsymbol{s}}_t = \boldsymbol{J}_{\boldsymbol{s}}\dot{\boldsymbol{\omega}} + \frac{\delta s}{\delta t}$$

Jacobian computation using finite differences. Each element of the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{s}}$ encodes a partial derivative of *visual motion features* values ($\boldsymbol{s}$) over each *warping unit* feature ($\boldsymbol{\omega}$):

$$\boldsymbol{J}_{\boldsymbol{s}} = \left(\frac{\delta s_q}{\delta \boldsymbol{\omega}_k}\right)_{q,k}, q \in [0..V], k \in [0..W] \tag{2.5}$$

where V is the total number of *visual motion features* and W is the total number of *warping units*. A forward evaluation enables us to compute a variation in visual motion features $\Delta\boldsymbol{s}$ from a variation in degrees of freedom $\Delta\theta$ (for small enough variations):

$$\Delta\boldsymbol{s} = \boldsymbol{J}_{\boldsymbol{s}}\Delta\boldsymbol{\theta} \tag{2.6}$$

To solve the problem, we therefore reverse Equation 2.6 by approximating J^{-1} using a damped least square method, and we obtain the variation of degree of freedom, that we expressed as variation on $\boldsymbol{\omega}$ (linear combination of $\boldsymbol{\theta}$, see Section 2.1.3).

$$\Delta\boldsymbol{\omega} = \boldsymbol{J}^{-1}\Delta\boldsymbol{s} \tag{2.7}$$

The input vector $\Delta\boldsymbol{s}$ is classically computed as the difference between the expected features and the measured features $e = s^* - s_t$ and capped with a maximum threshold:

$$\Delta\boldsymbol{s} = \begin{cases} e & if\ \|e\| \leq DS_{max} \\ DS_{max}\frac{e}{\|e\|} & otherwise \end{cases} \tag{2.8}$$

In practice the computation of the Jacobian $\boldsymbol{J}_{\boldsymbol{s}}$ at any time t requires to evaluate each *visual motion feature* for each $2 * W$ variation of warping unit parameters. This is performed using finite differences, where variations in the agent motion need to be rendered to assess visual features.

We finally build the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{s}}$ as:

$$\boldsymbol{J_s} = \left(\frac{s_q^{\boldsymbol{\omega}_k^+} - s_q^{\boldsymbol{\omega}_k^-}}{2\boldsymbol{\omega}_k} \right)_{q,k,t} \tag{2.9}$$

We then estimate the inverted Jacobian, and use it to extract the warping direction $\boldsymbol{\Delta\omega}$ (see Equation 2.7). An additional clamping is applied to the obtained vector to smooth the final motion modification. Finally, the new warped motion is computed using Equation 2.4. Implementation details and the setting of the expected feature s^*, for different case studies, are detailed in the following section followed by a user evaluation performed using Virtual Reality.

2.2 Results

We demonstrate our method over three case studies, with one virtual agent acting in front of one observer. We focus on upper-body nonverbal communication mainly using head, torso, arm and hand body parts, although the proposed method is suitable to control any body limb. Of environment conditions and related visual targets, we explore: i) the influence of changes in the observer's viewpoint, ii) the influence of occlusion or lighting conditions, and iii) the potential exaggeration of extraverted traits of a motion. Video examples of these three use cases and of additional example are available [1]. Results and performances are discussed in this section, the method is more generally discussed in Section 2.3.4.

2.2.1 Implementation

We implemented our technique using *Unreal Engine* both for versions 4 and 5. The approach runs at interactive frame-rates (>30fps on a computer equipped with an Intel Core i7-9850H CPU @ 2.60GHz, 32GB of RAM, Nvidia Quadro T2000) and can be used in interactive and non-interactive contexts. **Estimators** (see Section 2.1.2) were implemented using shaders, while the visual motion feature vector was computed through multiple rendering passes. The **warping operators** (see Section 2.1.3) were built above *Unreal* control rigs which provide a direct access to the degrees of freedom of the skeletal structure of animated characters. Our virtual agent is based on a *Meta-humans* model. The baseline motions (unwarped) were either recorded using a motion capture system (Xsens suit) or

1. https://youtu.be/8bQWD0WWnP4

Joint	Movement
Spine	Bending forward - backward
	Bending left - right
	Rotation around vertical axis
Neck-head	Bending forward - backward
	Bending left - right
	Rotation around vertical axis
Shoulders	Flexion - extension
	Abduction - adduction
	Internal - exterma; rotation
Elbow	Flexion - extension
	Rotation on its axis
Wrist	Flexion - extension
	Ulnar - radial deviation
	Supination - pronation

Table 2.1 – Defined pose warping units for upper-body motions. The last three are independent for each arm.

taken from a public database (Mixamo). During the execution, the **control loop** (see Section 2.1.4) generates $(2 * W) + 1$ copies of the virtual agent, at each frame, from the observer's point of view: 1 copy is used as a reference (unwarped motion) for visual motion features evaluation and for the target definition, whilst the $2W$ others are used to compute $\boldsymbol{J_s}$ by rendering warped motions, for each of the W warping units, in both warping directions.

To encode the different adapting behaviors, the designer first needs to decide the relevant *body parts* on which *visual motion features* are computed, then select the subset of these features to control. For each visual motion feature the designer can specify the magnitude and direction of its change over time, relative to the current value or to an absolute visual target. The magnitude affects mostly the responsiveness, and the direction defines the sign of the adaptation. Both values define the final visual target s^* (see Section 2.1.4). Consequently the designer needs to choose the set of active *warping units*, *i.e.* parts of the body which animations will be warped (see Table 2.1). We will describe and justify these choices in each example.

Figure 2.3 – Results relative to a viewpoint change, here toward the observer. On the left (case 1.1) we show three examples of increment in visual coverage for the highlighted body parts, while, on the right (case 1.2), an example of adjustment related to an object held in hand (a tablet here).

Case	Body parts	Visual motion features	Warping units
Case 1.1	face torso hands	*visual coverage* *visual extension*	spine neck-head arms
Case 1.2	face tablet	*visual coverage*	spine neck-head right arm

Table 2.2 – Selected values for Case 1.1 and Case 1.2.

2.2.2 Case studies

Case 1 - viewpoint changes. The purpose of this first study is to experiment the influence of viewpoint change on the animations to edit.

Scenario 1.1: In this scenario, the virtual agent aims at catching the attention of the observer while performing a two-hand waving motion. We recorded the baseline motion assuming that the agent was facing the observer. The objective is to modify this motion for a different camera angle, in a way that the agent better captures the observer's attention according to his/her position. We performed the visual evaluation with the face and the waving hands as body parts linked to the visual target. We selected *visual coverage* as the visual motion feature and specified a value to maximise frontal visual appearance. Also,

when positioning the observer at different distances from the camera, we used the control of the *apparent horizontal extension* to adapt the waving amplitude. The *warping units* related to the agent's spine, neck and arms were selected since they affect the waving motion of both hands.

Scenario 1.2: Here, the virtual agent shows to the observer an object held in its right hand. Again, our baseline animation was recorded with the observer facing the agent. The challenge is to adapt arm and hand motions to the observer's point of view in such a way that the object shown appears in the center of the observer's FoV. Regarding the relevant limbs related to the visual target evaluation, we selected the head and the additional external object (a tablet) held in the hand. Similarly to the previous scenario, we aimed for an increment of the *visual coverage* of the face and the tablet screen to maximise their visual appearance. The selected *warping units* were the ones related to the spine, neck and right arm of the agent since the motion was performed with this arm only.

Informal analysis: Scenarios 1.1 and 1.2 are illustrated in Figure 2.3 left. They were tested with different viewpoint parameters, namely, position, orientation, static/dynamic motions (see also the related **video**) [2]. Results show that our approach successfully adjusts motions to changing viewpoints. Indeed, the first scenario demonstrates how the waving amplitude adjusts according to the observer's distance. In the second scenario we show how a unitary motion could be rapidly warped, to fit the duration of the action, with the same target of maintaining the visibility of a relevant object shown in the observer's FoV. We also show that our approach allows for the control of multiple limbs by generating subtle variations on a motion without affecting its original purpose. One could argue that similar results, especially Scenario 1.2, could be replicated using an inverse kinematics (IK) approach. This is only partly true, as our approach based on visual motion features also integrates environmental conditions such as lighting or scene layout (*e.g.* bring an object in front of someone, in light) with the exact same setting.

Case 2 - occlusion/visibility. In this case study we aim at exploring through two scenarios the warping of agent limb motions to ensure proper visibility from the observer's viewpoint.

Scenario 2.1: In the first scenario, we recorded a one-hand waving motion, fully visible to a facing observer. The objective is to adapt this motion to improve its visibility by accounting for environment effects – occluding or partially occluding the animation. We

2. https://youtu.be/8bQWD0WWnP4

Figure 2.4 – Results regarding the occlusion/visibility use case. We demonstrate the influence of solid or sparse occluders in the control of visibility, both in static and dynamic conditions: increment for case 2.1, decrement for case 2.2. Visual coverage was used as the regulating visual motion feature.

Case	Body parts	Visual motion features	Warping units
Case 2.1	face upper torso right hand	*visual coverage*	spine neck-head right arm
Case 2.2	head torso	*visual coverage*	spine neck-head arms

Table 2.3 – Selected values for Case 2.1 and Case 2.2.

considered as relevant limbs the ones that usually help capturing the observer's attention at a distance *i.e.* the waving hand, the face and the upper torso. For the same reason, the *visual coverage* was selected as the active *visual motion feature*. In this case, an increment of this feature for the hand, face and upper torso would improve the perceived visibility of the motion. The selected *warping units* influenced the spine, the right arm and hand.

Scenario 2.2: Here, our agent tries to hide from the observer; to achieve this, we captured a crouching motion as if someone was hiding behind an object, and used it for the original animation. The objective here was to adapt the motion to make the agent less visible from the observer's viewpoint by hiding behind occluders. A decrement on *visual coverage* (the selected *visual motion feature*) was selected as target. Finally, we also tried to simultaneously increment the perceived *visual coverage* of a specific limb while hiding the rest, *e.g.* maintaining eyes visible. We selected as *warping units* those related

to the spine and the neck, and as relevant limbs all the upper body ones. Using similar parameters, we demonstrate how, in a different situation (see Figure 2.4, case 2.2 right), the agent would use an object (the shield) to get cover and reduce the visibility of the head.

Informal analysis: Scenarios 2.1 and 2.2 are illustrated in Figure 2.4 center. These scenarios test different visibility parameters: static/dynamic occluders, static/dynamic lighting, different kinds of obstacles (see also the related **video** [3]). Results show that our method successfully adapts motions to visibility conditions and targets. Indeed, scenario 2.1 demonstrates how a waving motion initially recorded in a clear view situation facing a frontal observer can be adapted in other visibility conditions, by bending and adjusting the configuration of the arm toward a space where it was more visible. Scenario 2.2 shows how a crouching motion can be adapted to increase its hiding purpose, and we also show that our method enables combining this purpose with additional minor behaviors such as maintaining the top of the head visible. Such results demonstrate the advantage of our visual approach for these kinds of situations, by enhancing and adapting motions in different visibility conditions.

Case 3 - expressivity. Here, our aim is to explore how visual features could be exploited to control the expressivity of a motion with our approach.

For the current example, we aim at influencing *extraversion*, one of the Big Five traits of personality [Goldberg, 1990]. This trait describes someone who typically captures the attention of an observer, is enthusiastic, energetic and sociable. In this scenario, the virtual agent is performing communication gestures as if it was talking. For the original animation we motion captured an actor conversing with the experimenter in a neutral way. The objective here is to adapt the arm and the hand motions to tune the trait extraversion of the agent, taking into account the observer's viewpoint. We selected *warping units* related to the head, elbow, shoulders, arms and hands, and the selected limbs were arms and hands. *Visual coverage, apparent vertical and horizontal extensions* were selected for the *visual motion features*. The visual target was then to increase/decrease both extensions to make the agent appear more/less extraverted. Indeed, Neff et al. [2010] described several modifications to the character's gestures with a perceptual effect of creating an extraverted personality – by increasing spatial scale of the strokes, elbow rotations outwards (arm swivel) and increasing the shoulder raise. On the opposite, by decreasing the

3. https://youtu.be/8bQWD0WWnP4

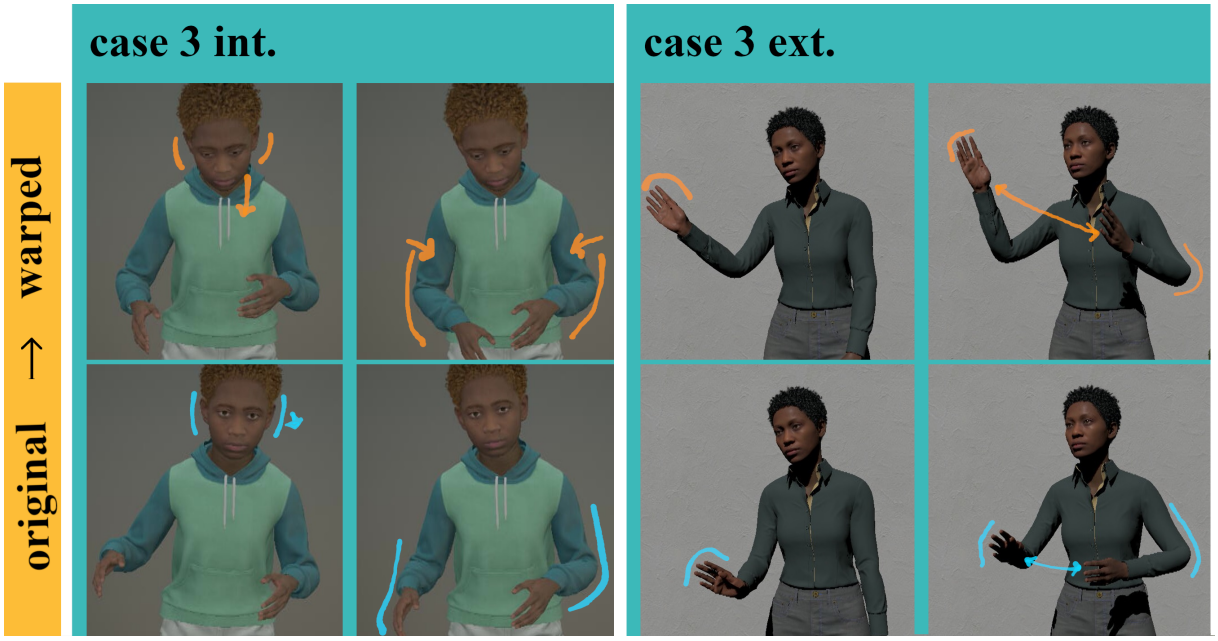


Figure 2.5 – Details regarding expressivity control with our approach. In the left block (case 3 int.), we highlight how the averted orientation of the face and the reduced openness of the arm motion produce a more introverted behavior. Oppositely, on the right block (case 3 ext.), the horizontal extension of the posture and the wider amplitude of the gesture simulate a more extraverted appearance.

Case	Body parts	Visual motion features	Warping units
Case 3	face torso arms hands	*visual coverage* *visual extension*	spine neck-head arms

Table 2.4 – Selected values for Case 3 introvert and Case 3 extravert.

same characteristics, we define an introvert personality. Additionally, for this case, we have controlled *Visual coverage* of the head to control the head posture.

Informal analysis: Scenario 3 is illustrated in Figure 2.5 (see also the related **video** [4]). In this case study, we show that our method can modify gestures in similar ways than previous studies [2010], which should also modify the perceived introversion/extraversion of the character.

4. https://youtu.be/8bQWD0WWnP4

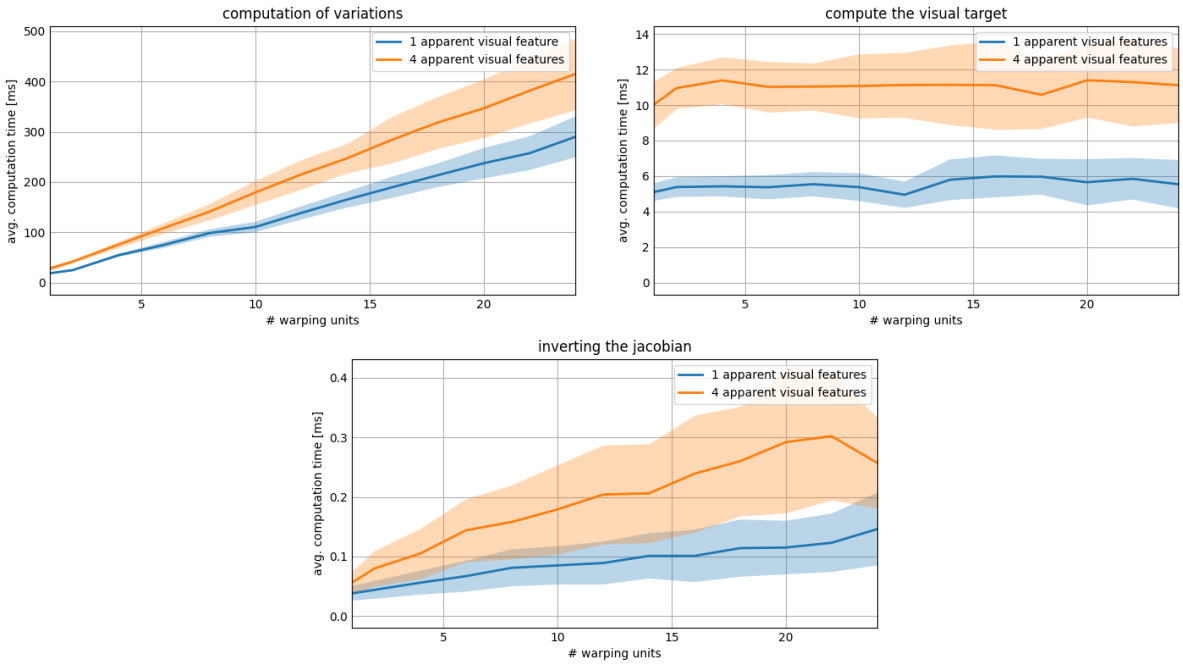


Figure 2.6 – Average computation time for three main parts of our technique, depending on the selected number of *warping units*. Left: time to compute the variations of *visual feature* values. Middle: time to compute the visual target in relation with the current *visual feature* values. Right: time to invert the Jacobian matrix. Results consider either one (blue) or four (orange) active apparent visual features. Performance evaluated on a computer equipped with an Intel Core i7-9850H CPU @ 2.60GHz, 32GB of RAM, Nvidia Quadro T2000.

2.2.3 Performance

To objectively evaluate the computational load of our approach, we also present some information about the computational resources required for the application of our method. The core computation runs on a dedicated thread at a frequency set between 3 Hz and 5 Hz, according to the selected parameters (*e.g.* warping units and visual features number). This thread is provided with the rendered scene and variations, and outputs the warping parameters for the virtual character. At each frame, the separated rendering thread updates the posture of the character following the warping result. This separation of threads allows the code to achieve real-time performance and work at an interactive framerate in virtual reality, even though the proper computation is not updated every single frame.

Figure 2.6 shows the effects of the number of selected *warping units* (horizontal axis) and *apparent visual features* (blue and orange curves) on computation time. Values were

captured on a machine equipped with Intel Core i7-9850H CPU @ 2.60GHz, 32GB of RAM, Nvidia Quadro T2000. *Warping units* size has a stronger impact on performance for the computation of variations (order of magnitude $10^2 ms$ in Figure 2.6 left), due to the need of generating and evaluating each variation of each warping unit. Nevertheless, the computation time is impacted linearly with the size of *warping units*. However, this heavy computation also further justifies the definition of *warping units* as a summary of the structure's degrees of freedoms, instead of computing variations on individual joints separately. On the other hand, the estimation of the *visual feature* values for the character, from which we extract the visual target (Figure 2.6 center), depends only on how many *apparent visual features* we consider in the computation. Finally, we show that the time to invert the Jacobian depends linearly on its dimensions VxW – number of *visual features* per number of *warping units* – and has minor impact on computation time (order of magnitude $10^{-1} ms$ on Figure 2.6 right). In the presented implementation, the maximum possible size for the Jacobian matrix is 4x22, but none of the presented case required more than half of these dimensions.

2.3 Evaluation

Since our main objective is to design reactive virtual humans who are able to interact and initiate an interaction with users, we designed an experiment in VR, where a virtual agent attempts to catch the user's attention by waving at him/her (similarly to case 1 and 2 from Section 2.2.2). In such a scenario, we aim at evaluating whether participants are able to detect if a waving motion is directed toward them or toward another agent.

We also compare our technique to a standard approach based on orientating the root of the character. We hypothesize that our technique will outperform the simpler approach, showing a higher success rate to detect the target of the waving motion, as well as maintaining or increasing the naturalness of the interaction.

2.3.1 Procedure

Participants. Sixteen unpaid participants volunteered for the experiment (3F, 13M; age: avg=25$\pm$3, min=20, max=30). They were all naive to the purpose of the experiment, had normal or corrected-to-normal vision, and gave written and informed consent. The study conformed to the declaration of Helsinki, and was approved by the local ethical

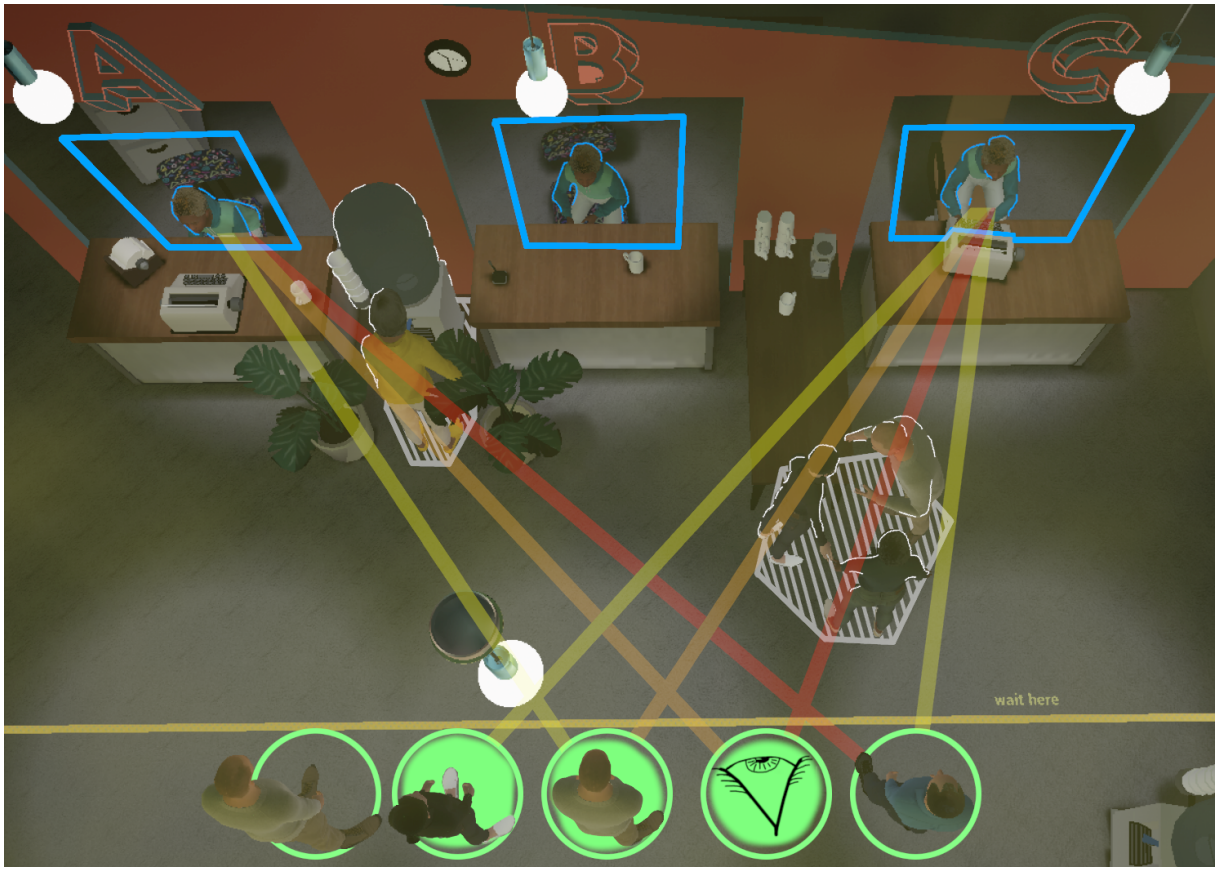


Figure 2.7 – The disposition of the agents in the virtual scene: starting from the bottom (green circles) the position of five customers, including the three (filled green circles) where the participant could be spawned. In the middle (highlighted in white) is the main area of visual occlusion, composed of objects and virtual agents. On the top: the three desks, each with a virtual clerk (blue squares). Yellow, orange and red lines denote the degree of partial occlusion in the combined view direction between customers and clerks (yellow for low degree of occlusion, orange for medium, and red for high).

committee (COERLE). Participants were first asked to read and fill up a consent form. They were equipped with a HTC Vive's headset, headphones and two controllers. The experiment was performed in a standing position and participants were able to take a break between each trial if needed.

Task. The evaluation takes place in a virtual office, as presented in Figure 2.7. The participant is spawned in a line of virtual customers, who are waiting to be called by one of the virtual office clerks, using a hand waving motion. In total there are three clerks, each one sitting behind a desk. In between the customers and the clerks are additional agents and static objects that occlude the view of some of the clerks.

Figure 2.8 – An example of the point of view of the user during the interaction. In the displayed situation, the virtual clerk, in the central desk (blue square), is waving toward a virtual agent positioned on the left of the user (the blue arrow indicates the direction of waving). In this case the field of vision of the user is free of occlusions. The blue square and arrow are displayed for informative purpose, and were not shown to particpants.

The participant stands in one of the three central positions of the customers group, and is exposed to one of the following interactive situations, where one of the clerk waves in 3 possible directions: the participant (i), one of the other virtual customers to the left (ii) or to the right (iii) of the participant. The motion of the clerk is animated using two techniques: a straightforward orientation of the whole body of the agent toward the target (SF_adaptation), and our technique (visualWarping) parameterized with an increment in the *apparent visual coverage* of the head and the waving hand, using all the *warping units* affecting these two body parts. Additionally, for both techniques, we implemented the same gaze controller to orient the eyes of the agent toward the selected customer.

In total, participants performed 2 training trials and 54 experimental trials presented in a randomized order (2 techniques $\times$ 3 directions $\times$ 9 repetitions). After each trial, participants were asked to: (i) answer the following question "According to you, to whom was the interaction directed toward?" with the possibility of selecting one of the customer's position, including their own, and (ii) to rate the level of realism of the interaction on a

scale from 1 (low) and 9 (high). At the end of the experiment, we also collected answers using a post-experiment questionnaire composed of questions related to demographics (gender, age and familiarity with VR), the strategy used ("What strategies did you use to identify to whom the interaction in the virtual reality experience was directed to?") as well as self-reported free comments.

2.3.2 Analysis

For each participant and each condition, we computed the answer accuracy (i.e., the ratio of correct answers across repetitions), as well as the level of realism (the average level across repetitions). To evaluate the effect of the technique while considering the effect of the interaction direction of the waving motion, we conducted a two-way repeated measures Aligned Rank Transform (ART) Anova. When relevant, we performed Tukey post hoc pairwise comparisons. Statistics were performed using R and the level of significance was set to 0.05. Results are presented using mean±SD.

2.3.3 Results

Results showed a significant main effect of the technique on accuracy answers ($F(1,15)=7.09$, $p=0.017$, $\eta_p^2=0.32$), with better accuracy with visualWarping (0.67 ± 0.2) than with SF_adaptation technique (0.59 ± 0.25), as illustrated in Figure 2.9, left. There was also a main effect of interaction direction ($F(2,30)=18.92$, $p<0.001$, $\eta_p^2=0.356$), where participants were less accurate when the virtual human was interacting to the right of the participants (Figure 2.9, right).

There was no significant interaction effect ($F(2,30)=2.19$, $p=0.12$, $\eta_p^2=0.12$), but, as illustrated in Figure 2.10 left, there is a tendency for our method to be more efficient than SF_adaptation in direct and right relative interaction direction.

Regarding the perceived level of realism, results showed only a main effect of the interaction direction ($F(30,2)=3.97$, $p=0.03$, $\eta_p^2=0.21$), with a lower realism when the virtual human is interacting to the right than toward the participant (Figure 2.10 right).

2.3.4 Discussion

Overall, the presented results highlight the impact that vision-based editing techniques might have on human-agent interaction in Virtual Reality. Even though we have

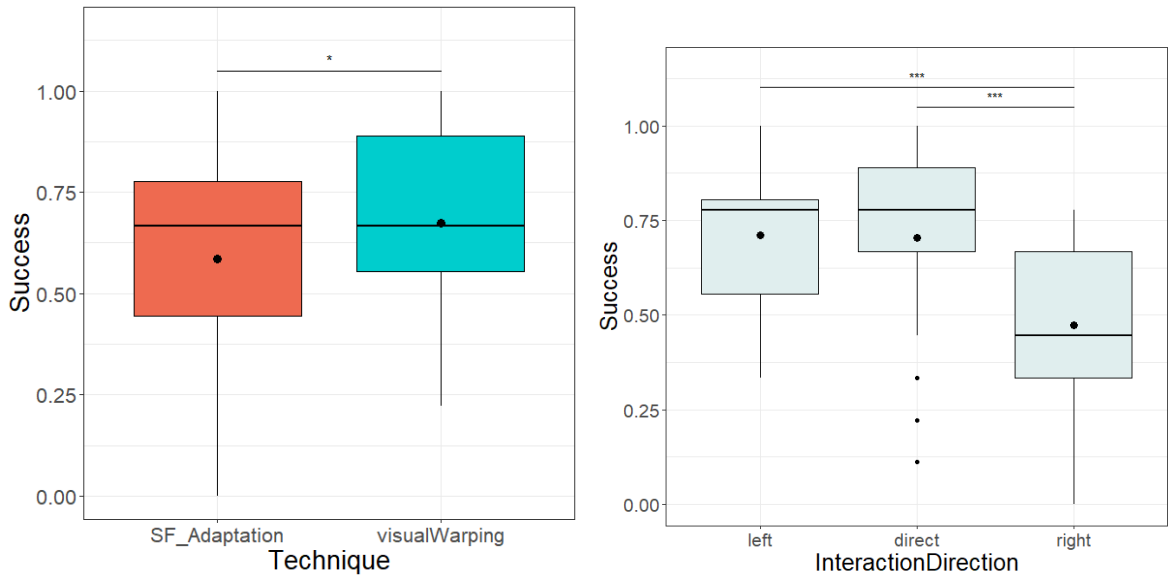


Figure 2.9 – Main effects of the Technique and Interaction direction on participants' answer accuracy.

not detected an improvement in the perceived realism of the interaction, meaning that participants rated our edited interaction as realistic as the oriented one, we observed a pertinent increase in the accuracy. Thus, participants could interpret easily the agent's aim when adapted with our technique in comparison with using a more simple one. In apparent contrast with the current discussion, all the participants mentioned eye behavior as one of the main and sometimes the only (25%) feature used to identify the interaction target. In fact this is not surprising, and even though eyes play a crucial role in user-agent interactions [Beebe, 1974; Kleinke, 1986], they are not always reliable, especially in situations when the proximity between the potential targets is narrow, or the eyes are not clearly visible (*e.g.* when the agent is far or not oriented toward the player). Accordingly, in the next chapter, we will present a study on how gaze behavior can affect the perception of virtual agents. Finally, the results suggest that our technique can control body non-verbal communication, mentioned as secondary or primary feature in 75% of the answers, to enhance the interpretation of the waving orientation.

Moreover, it is important to discuss the disproportion we obtained between different relative directions of the target, for both techniques. We believe that this effect is caused by the design of the scene. Indeed, the right-most customers positions present, in general, a higher degree of occlusion relative to the clerks' desks, than the left ones (see Figure 2.7). Under this circumstance, we reported a negative effect, in realism and accuracy of both

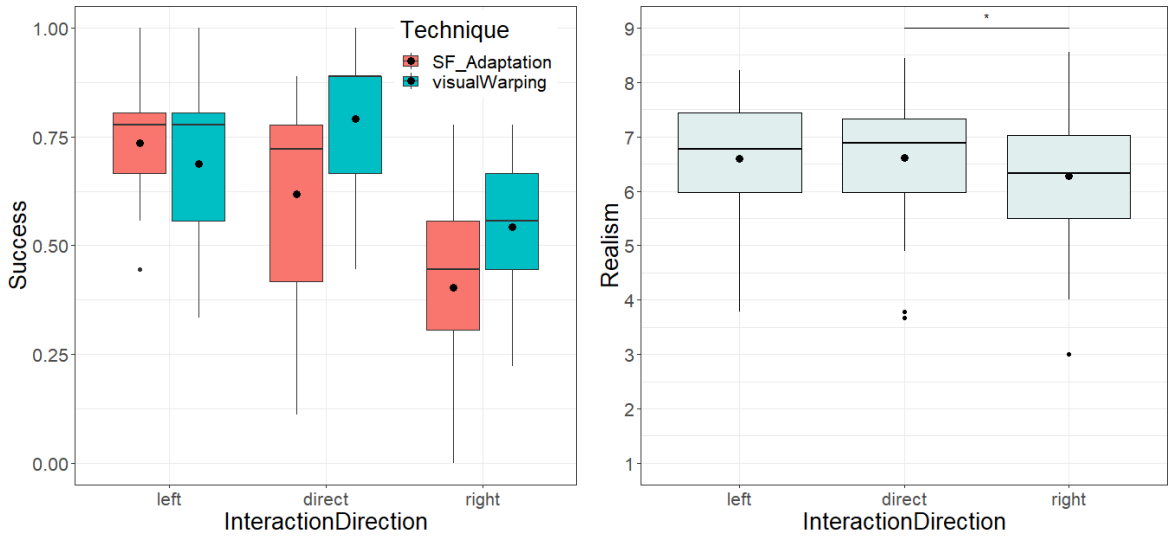


Figure 2.10 – Left: participants' accuracy answers depending on the technique and interaction direction. Right: main effect of interaction direction on the perceived realism.

techniques, for the relative right targets, which is partially mitigated in the accuracy by our technique.

2.4 Conclusion

This contribution proposes a motion warping technique that enables linking low-level motion variables with FoV-dependent visual motion features. Our results show that our approach is effective and applies to various cases, such as adjusting motions to changes in an observer's position, environment or lighting conditions. We also believe that this technique can be exploited to influence the expressions conveyed by animations (*e.g.* intro vs. extraversion) thereby helping designers to fine-tune the personalities of their characters or having virtual characters adapt their non-verbal communication toward observers (or avatars). Our proof of concept and the user experimentation in Virtual Reality validate the visual servoing scheme and yield a general and promising solution. While the method presents some analogies with inverse kinematics methods, through-the-lens techniques, or line of action control, it offers higher levels of control than just positions and velocities of joints in the image space.

Limitations and future works. Currently, our method is limited in multiple ways. First, it requires a prior selection of visual motion features to be controlled, warping units

to activate, or magnitude and direction of visual features. Methods to select relevant combinations of parameters would improve the practical usability of the approach. Inappropriately chosen parameters can generate unwanted results that are partially filtered out by joint limits and stiffness. Additional kinematic filters of motion editing (*e.g.* balancing the center of mass, remaining in the human motion manifold, preventing self-collisions) could help reduce the negative effects of parameter tuning. In addition, we only explored a limited set of spatial visual features while operators could also perform time warping, and measure features over a sliding window rather than on a per-frame basis. Directly integrating computational saliency techniques [Bruce et al., 2015] closer to visual attention mechanisms could also help to guide the warping of animations. Additional saliency biases could be added to account for top-down attention mechanisms specific to characters (*e.g.* focus of attention on head, eyes and hand movements) to build an attention-driven approach.

Second, the design of our warping units remains empirical. Existing work to automatically define rigging functions [Holden et al., 2016b] could improve over our solution. In our case, we could explore means to automatically correlate low-level motion variables with the variations of visual motion features, with the difficulty that these relations depend on the motion performed, the desired goal of editing, and the observer's position.

At this stage, we also left apart the question of setting the *appropriate* levels of visual motion feature editing. We partially address this question for specific use cases, *i.e.* the attention-catching scenario where the levels are set as if observed from an optimal angle, defined during motion capture. Still it remains challenging in other situations, in particular in the case of controlling motion expressivity. By which level a feature should be adapted to change the expression of motion? Which parameters or kinematic limitations enable reaching the desired editing levels? A data-driven approach would be relevant to address these questions. The difficulty is twofold: one is to gather the required amount of data to capture feature level variations with corresponding semantics, the other one is to deal with human variability in such behaviors.

In this work, we focus on upper-body motion, with the intention of controlling gestures that are salient for non-verbal interaction. However, we believe that the current paradigm, based on controlling high-level visual features, can be applied to a heterogeneous set of motion synthesis tasks: lower-body motion, *i.e.* steps generation, or action planner, *i.e.* which hand to use in the waving motion, or when to trigger an action. Indeed, in a previously published short paper [Raimbaud et al., 2021], we introduced the concept

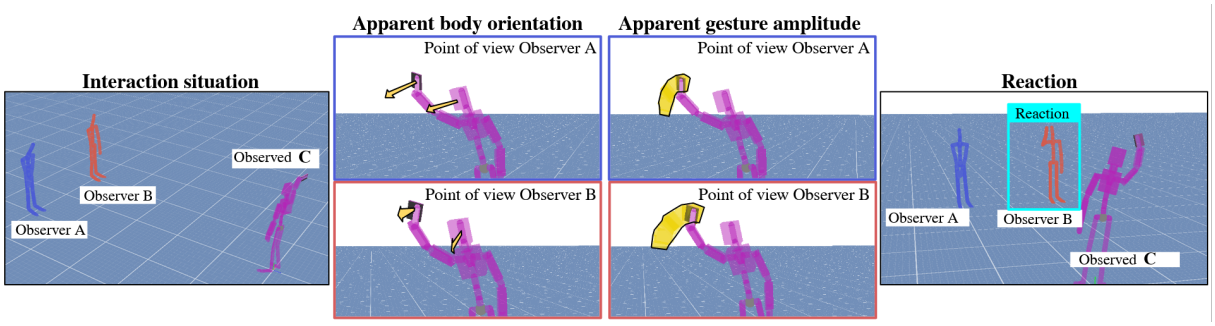


Figure 2.11 – Waving case from left to right: two virtual humans – A and B – observe another one. Visual motion features – body orientation, gesture amplitude – are computed on the motion of the observed virtual human (C) from each viewpoint. Accordingly, to the computed features, we identify that the interaction is directed toward the virtual human B and the reaction is therefore triggered.

of analysing visual features to help trigger proper character reactions (Figure 2.11). We believe that such synthesis tasks would help improve the realism of the interaction.

In our examples, we explore warping the motion of a single character. Undoubtedly, it is relevant to discuss the case in which multiple characters are warped simultaneously. Considering a group of characters seen from a single point of view, our approach would adapt their configuration to optimize a joint set of *apparent visual features* for that point of view. In terms of performances, this would require the evaluation of variations for each single character, but, as in Section 2.2.3, the total consumption would depend mostly on the total joint number of selected *warping units*. Rendering would however not require additional computation, except the necessary meshes and textures. The dual case considers multiple point of views for a single character. Even though this approach is feasible, it would require the management of a compromise between possibly contradicting visual targets.

Conclusion. In conclusion, the technique we propose is a viewpoint dependent motion editing approach that exploits a number of visual features from an external observer's point of view to drive and warp animations. As a result, this can empower creative artists, but also autonomous characters with means to control the information they convey to an observer.

Chapter 3

Virtual character gaze in virtual reality: exploring the stare-in-the-crowd effect

Contents

In this chapter, we present a user study on which I collaborated awith Pierre Raimbaud during my PhD. My contribution on this project was on the conceptualization, development, investigation and partially the writing of the research pubblication. In continuation with the previous chapter, where we evaluate the effects of edited upper-body motions on the understanding of the agent intentions, in this chapter, we expand the study of non-verbal interactions exploring the effect gaze of behavior on the observer's

Figure 3.1 – The stare-in-the-crowd effect describes the tendency of humans in noticing and observing, more frequently and for longer time, gazes oriented toward them (directed gaze) than gazes directed elsewhere (averted gaze). This work analyzes the presence of such an effect in virtual reality and its relationship with social anxiety levels. The figure above shows an example of the user's view during our experiment. All agents, except the woman in the front row wearing a black jacket, have their gaze averted.

perception.

We previously mentioned that non-verbal cues are crucial in real-world interactions, and, in particular, we saw that gaze has a primal role in these interactions and in the observer's perception(see Section 2.3). In this work, we focus on the initiation of an interaction between virtual humans and a user [Mohammad and Nishida, 2008], and we ask whether the virtual humans' gaze behavior can be useful in initiating it. This phenomena have been addressed through a protocol called the **"stare-in-the-crowd effect"** [Von Grünau and Anston, 1995], which demonstrated that when multiple faces are exposed to a subject during a visual search task, the detection of the ones whose gaze is **directed** towards the subject is faster than the ones whose gaze is not looking at him/her (**averted**). It has also been shown that in free visual tasks, visual attention is affected by the presence of directed gaze among averted ones [Crehan and Althoff, 2015].

The **stare-in-the-crowd effect** is a gaze behavior effect that reflects the existence of a search asymmetry between directed and averted gazes when users face a crowd: directed gazes are detected faster than averted ones and cause more frequent and longer fixations [Von Grünau and Anston, 1995]. Previous works proved the effect using various stimuli, *e.g.* photographic [Crehan and Althoff, 2015; Doi and Ueda, 2007; Ramamoorthy et al., 2019; Von Grünau and Anston, 1995] and 3D geometric representations of faces [Colombatto et al., 2020], and conclude that the effect is not necessarily due to a particular saliency of the eyes but rather due to the intentionality conveyed by the

Figure 3.2 – An example of the visual stimuli by Crehan et al. [2015].

stimulation.

It should also be mentioned that some studies have mitigated the existence of the stare-in-the-crowd effect, notably by refuting the fact that this effect occurs in every configuration [Cooper et al., 2013; Palanica and Itier, 2011a; Palanica and Itier, 2011b]. In particular, the usual search task commonly used is criticized by Crehan et al. [2015]. Crehan et al.'s study [2015], therefore, proposes a new evaluation paradigm: based on an observation task, while measuring the effect through eye-tracking. With this paradigm and using photographs with complete bodies they also observed the stare-in-the-crowd effect. Moreover, they also included dynamic conditions, where gazes changed from averted to direct ones and vice-versa. Such dynamic conditions replicate some natural eye-gaze interactions, such as *being caught staring at someone* and *catching someone else staring*. They found that these dynamic conditions affect user gaze behavior similarly to directed gazes.

In this study, we investigate whether the stare-in-the crowd effect is preserved in Virtual Reality, replicating the experiment of Crehan et al. [2015] (see Figure 3.1 and Figure 3.2). To this end, we designed a within-subject experiment where we analyze 30 human users' gaze behavior when observing an audience of 11 virtual agents following 4 different gaze behaviors. We computed fixations and dwell time, and we also collected the users' social anxiety score using a post-experiment questionnaire to control for some potential influencing factors. Results show that the stare-in-the-crowd effect is preserved in virtual reality, as demonstrated by the significant differences between gaze behaviors. Additionally, we found a negative correlation between dwell time towards directed gazes and users' social anxiety scores.

This chapter is structured as follow. In Section 3.1, we introduce the objective and hypotheses about our virtual reality replication of the stare-in-the-crowd effect. Section 3.2 details the experiment, and Section 3.3 our results. We finally discuss them in Section 3.4 before concluding in Section 3.5. Additional results are included in the related Appendix A.

3.1 Objective and hypotheses

In the present study, we aim at investigating the perception of non-verbal cues when a user is immersed in a virtual environment populated with virtual agents. Our main objective is to study the reaction of users, through their gaze behavior, when facing a virtual crowd where agents can either look at them or look away. Previous studies using eye-tracking investigated user's gaze when observing photographs depicting a seated audience. They showed users' preference for gazing at individual subjects in these photographs, whose gaze was directed towards them rather than averted from them, also called the stare-in-the-crowd effect [Von Grünau and Anston, 1995]. According to the literature, virtual reality can be used to depict social interactions with user's behaviors that are close to real-life ones. We are thus interested in the presence of this effect in virtual reality – an environment more adapted to natural human interactions than photographs.

Hypotheses. Towards this objective, we propose two hypotheses, H1 and H2. First, we expect that we will observe the same effect as reported in Crehan et al. [2015] using a series of photographs, but in virtual reality.

— **H1**: The stare-in-the-crowd effect is preserved with virtual agents in virtual reality. This means that eye-tracking data will show more salient characteristics (number of fixations, gaze duration) towards the agent who is directing its gaze towards the user, as opposed to when the agent is not looking at the user. Moreover, we also expect the same effect comparing the static averted condition to each dynamic one, *i.e.*, during the phenomena *being caught staring* and *catching someone else staring*. However, for these gazing conditions we expect a lower effect magnitude than for the static directed gaze one, since the time when the agent is looking at the user is shorter. Finally, we are also interested in the comparison between the behavior of the user in the dynamic conditions as opposed to the static directed one.

Moreover, it has been shown previously that social anxiety influences virtual reality users' gaze behaviors towards a virtual crowd, in a similar way to when interacting with

humans in physical reality [Lange and Pauli, 2019; Wieser et al., 2010]. Indeed, a higher social anxiety is typically correlated with a lower rate of mutual eye contact towards directed gazes than in the case of socially non-anxious individuals [Baker and Edelmann, 2002; Schulze et al., 2013]. Therefore, we expect that:

— **H2**: There will be a negative correlation between the time spent gazing towards the agents who are staring at the user and the user's level of social anxiety.

This suggests a possibility that the stare-in-the-crowd effect will depend on the amount of socially anxious individuals in our test sample. With many users scoring high on social anxiety this effect could disappear completely, thus, it is relevant to explore this relationship. It is also important to note that in some cases a lack of gaze towards a socially anxious individual can be more frightening, as it can signal disinterest. However, we created the experimental conditions where the context of the averted gaze would not be interpreted like this.

3.2 Experiment

3.2.1 Overview

To study the stare-in-the crowd effect in virtual reality, we designed an experiment inspired by Crehan et al. [2015], which demonstrated the presence of this effect using photographs. In our experiment, the user is asked to observe a virtual crowd where the gaze of the virtual agents is manipulated according to a series of target conditions/behaviors, similarly to Crehan et al. [2015]. These crowd gaze conditions are:

— Averted - **A**: no virtual agent looks towards the human user during the observation task (see Figure 3.3 Left.1);

— Directed - **D**: one virtual agent, referred to as the "active agent", stares at the user at the beginning of the observation task and will keep staring at him or her until the end of the task, while no other virtual agent stares at the user (see Figure 3.3 Left.2);

— Averted-then-Directed - **AD**: no virtual agent looks towards the user at the beginning of the observation task, but the active agent will start staring at the user once looked at and will continue to stare until the end of the task (see Figure 3.3 Left.3);

— Directed-then-Averted - **DA**: the active agent stares at the user at the beginning

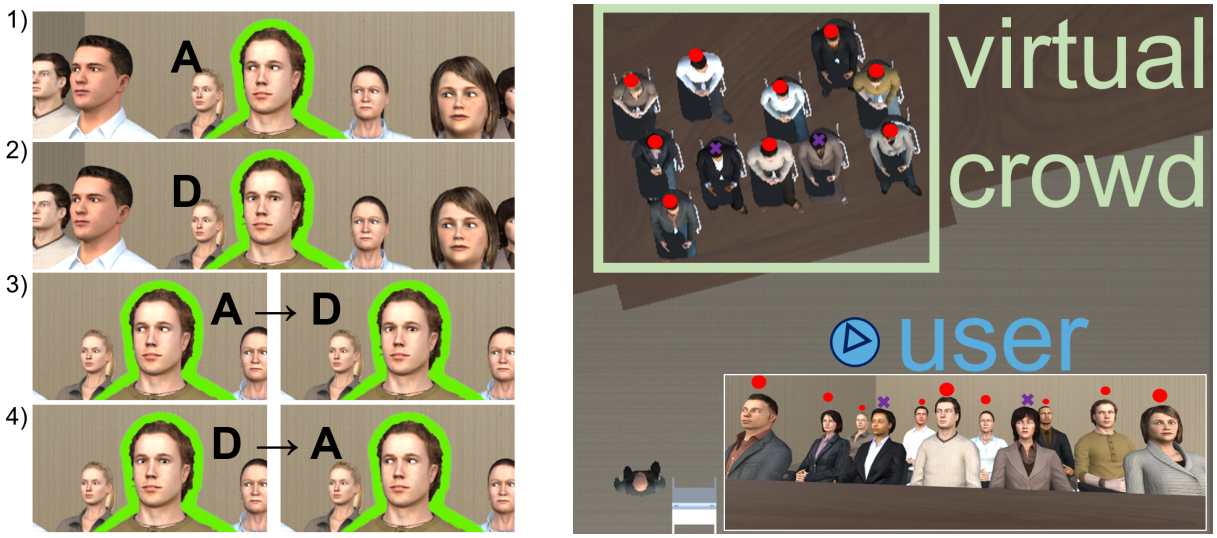


Figure 3.3 – Left: our four crowd gaze conditions (active agent in green): 1) averted gaze - **A**, 2) directed gaze - **D**, 3) averted-then-directed gaze - **AD**, and 4) directed-then-averted gaze - **DA**. See details in Section 3.2.1. Right: virtual scene where the user faces eleven agents listening to a speaker standing behind the user. The inset shows the user's view point during the observation task. Only active agents (red dots) are used to display a staring activity, to balance their distribution in the user's field of view as suggested in [Doi and Ueda, 2007; Palanica and Itier, 2011a; Palanica and Itier, 2011b].

of the observation task, but will stop once looked at, while no other virtual agent stare at the user (see Figure 3.3 Left.4).

Examples of such gaze behaviors in our virtual reality implementation can be seen in the supplementary video [1].

We asked users to observe the virtual crowd, without telling them to actively search for directed or averted gazes. Such indications are different with respect to some previous studies [Colombatto et al., 2020; Doi and Ueda, 2007; Ramamoorthy et al., 2019], but consistent with Crehan et al. [2015; 2021]. In line with Crehan et al. [2015], we also propose to use an eye-tracking system to evaluate the users' gaze behaviors instead of using a search task, which would be less natural. However, opposite to previous studies [Colombatto et al., 2020; Cooper et al., 2013; Crehan and Althoff, 2015; 2021; Doi and Ueda, 2007; Framorando et al., 2016; Ramamoorthy et al., 2019], we use a crowd of virtual agents in virtual reality as visual stimuli (see Figure 3.3 Right).

1. https://youtu.be/Ag3JPplVQdg

3.2.2 Virtual environment and stimuli creation

The virtual environment used, shown in Figure 3.3 Right, was created using Unity 2021.2.0b9. It is composed of a room, resembling a classroom or a conference room, equipped with standard pieces of furniture as well as individual chairs placed on a wooden stage. Virtual agents (our virtual crowd) are seated on these chairs, like an audience, $1m$ away from the user at the minimum. All virtual agents are clearly visible to the user, without any occlusion between their heads. The wooden stage hides part of the virtual agents' bodies, so as to make the user focus on their faces. Similarly to the photographic stimuli used in Crehan et al. [2015] (see Figure 3.2), the virtual audience was slightly (10°) oriented to the right, as well as the user (20°). Moreover, the user was placed slightly on the right of the virtual crowd. Such position/orientation choice was chosen for two main reasons: (i) to have all the virtual characters in the user's initial field of view, since they appear at real scale (1:1); and (ii) to allow virtual agents to look towards the user's position without needing to rotate their head, but only their eyes, while maintaining a natural gaze behavior (*e.g.*, horizontally rotating the eyes a maximum of 30° with respect to the head). These two aspects ensured that all virtual agents could be easily viewed, and that eyes orientation would be the main difference between them, with different gaze behaviors but similar head orientation, thus avoiding bias on these aspects [Marschner et al., 2015].

We used eleven virtual agent models from the Microsoft RocketBox adult avatars collection [Gonzalez-Franco et al., 2020], including six females and five males. Figure 3.3 Right shows this virtual audience from top and from the user's point of view. Additionally, we placed another male model in front of the crowd, as if he was giving a presentation to them. However, no speech could be heard by the user, it was only to provide a social setting, and to justify why the crowd was looking towards a common point away from the user. To increase the naturalness of agents' behaviors, we applied simple blinking animation on their eyes. Then, a specific gaze behavior was chosen according to the condition at hand, A, D, AD, or DA, as described in Section 3.2.1.

The virtual agent staring at the user, referred to as the "active agent", is chosen randomly among nine of the eleven agents of the crowd. These nine agents are highlighted with red dots in Figure 3.3 Right. This choice was driven by the need to have a balanced distribution of active gazing agents across the user's field of view, as suggested in [Doi and Ueda, 2007; Palanica and Itier, 2011a; Palanica and Itier, 2011b] to test any potential position effects on the results. It should be noted that for coherence with the other

conditions and to enable a consistent comparison of our metrics (see Section 3.2.6), an active agent is also chosen in condition A (no agent looks at the user), although it does not behave differently to the rest of the crowd.

Regarding agents' gaze behaviors, we proposed here to use a gaze model that favours eye rotations over head and torso rotations (while still providing realistic results), by applying rules, such as a maximum eye rotation angle of 30°, based on literature results [Gonzalez-Franco and Chou, 2014; Marschner et al., 2015]. Finally, in conditions AD and DA, where the agent's dynamic gaze behavior (from averted to directed or vice-versa) is triggered by the user, we introduced a time limit as suggested by Crehan et al. [2015]. If the user has not looked at the target agent within half of the total trial time, the agent's gaze changes anyway, without waiting for the user to look at the agent. Similarly, each trial repetition (*i.e.*, the user looking at the crowd) lasted 16 seconds. After this time, the environment fades out and fades in again to the same scene but featuring a new gazing behavior and active agent (see Section 3.2.5).

3.2.3 Participants and apparatus

Thirty participants (8 females, 22 males; age: average 30, Standard Deviation: 9.5; virtual reality experience from 1 to 5: average 3.4, Standard Deviation: 1.4; computer games experience from 1 to 5: average 3.5, Standard Deviation: 1.5) took part in our experiment, all with normal or corrected-to-normal vision. They voluntarily participated in the experiment and received no compensation for it. The study complied with the Declaration of Helsinki and was approved by local ethical committee (COERLE). Participants were asked to seat on a standard chair throughout the whole experiment, and to wear the virtual reality head-mounted display FOVE, which has an embedded eye-tracking system. Its field of view is 100°, both for visualizing the 3D scene and eye-tracking. The eye-tracker advertised spatial tracking accuracy is less than 1°, and its maximum sampling rate is 120 Hz.

3.2.4 Data collection

We collected two types of data: (i) continuous user's gaze behavior during the virtual reality experience, and (ii) social anxiety data afterwards.

Gaze behavior was collected using the embedded eye-tracking system of the virtual

reality headset. At each frame, the user's gaze information was logged along with the timestamp and the current gaze condition of the virtual crowd (A, D, AD, or DA). This gaze information was indicating the presence or the absence of a hit on the head of the "active agent", computed using the 2D screen position of the virtual reality user's gaze and the current 2D scene viewed by the user.

Information about users' social anxiety was collected after the experiment using the standardised questionnaire based on the Liebowitz Social Anxiety Scale [1987]. It enables the evaluation of social anxiety through self-estimation of the levels of fear and avoidance of a person in determined social situations. A score can be computed from the answers, ranging from 0 (not socially anxious) to 144 (very socially anxious).

3.2.5 Experimental procedure

First, an informative document about the study was given to the users, along with the informed consent form and oral explanations to answer any questions. Once ready, users were seated on a chair and equipped with the FOVE headset. A calibration of the eye-tracking system was performed to ensure the quality of gaze data collection.

Then, users were immersed in our virtual environment for a brief training phase, where they had time to familiarise themselves with the environment and setup. During this phase, all agents of the virtual crowd were looking at the virtual speaker, were not changing their gazing behavior over time, and random agents would be blinking in the crowd. Users were free to look both at the crowd and behind them to see the virtual speaker – which was not talking, to understand the context of the scene. It was explained to them that their task would be to face and observe the virtual audience, and to not look at the virtual speaker after the training phase. No information about gazing behaviors or any other specific task to complete were provided.

After this training phase, users were asked to perform 72 trials of this observation task, each lasting 16 seconds. All users were exposed to the same trials *i.e.*, all the tested conditions described in Section 3.2.1. Each combination of "gaze condition/behavior" per "active virtual agent positioning" was shown twice to each user, leading to: 4 gaze behaviors $\times$ 9 possible active agents $\times$ 2 repetitions = 72 trials in total. In order to make it possible for the user to rest during the experiment, the trials were ordered in 3 blocks, with equal number of gaze conditions presented in each block of 24 trials, as well as the distribution of the active virtual agent. The order of active agents was randomised

Figure 3.4 – Left: participant during a trial. Right: representation of the virtual crowd from the participant perspective.

inside each block. In averted conditions, an agent was chosen randomly and the position of these agents was balanced with the agents in the other conditions, which all include a directed gaze. Additionally, virtual agent models were randomly switched between all eleven positions, so that the appearance of the models would not influence the results. A 3-seconds black screen was displayed to the users between each trial. During this pause, users were asked to re-position their head and gaze orientation towards the top-center of the screen, by looking at a small geometric shape. This was done to ensure the same initial point for the user's gaze at each trial. Users were notified that the trials would be divided into three blocks of 24, so as to allow them to rest and remove the headset between each block to minimise fatigue. In addition, such breaks were also used to re-calibrate the eye-tracking system to ensure data quality. If needed, users could also stop within a block.

Finally, users were asked to fill a post-experiment questionnaire with the social anxiety questions, along with demographic ones (age, gender, experience with virtual reality and games) and a free comment section.

3.2.6 Metrics

From the eye-tracking collected data, we computed different metrics related to the users' gaze towards the active agent of the crowd. Gaze activity was split between saccades when such activity was shorter than 150 ms, and fixations when it was longer [Manor and Gordon, 2003; Westheimer, 1954]. For each trial, we considered the following metrics in line with Crehan et al. [2015]

— Dwell time: the total time in ms spent looking at the active virtual agent;

— Fixation count: the total number of fixations on the active virtual agent;
— First fixation time: the time of the first fixation on the active virtual agent, counted from the beginning of the trial;
— First fixation duration: the duration in ms of the first fixation;
— Second fixation time: the time of the second fixation on the active virtual agent, counted from the beginning of the trial;
— Second fixation duration: the duration in ms of the second fixation.

All the above metrics are used to identify the stare-in-the-crowd effect, particularly the dwell time and fixation count metrics that are computed even in absence of multiple fixations on the active agent. The analyses of first and second fixations are also important to better understand user's gaze behaviors. Even though they are not always present in stare-in-the-crowd related studies, they are particularly relevant for the dynamic conditions that we included here, where the user's first fixation on the active agent triggers the change in its gaze behavior (from averted to directed or vice-versa).

3.3 Results and discussion

In this section we detail the principal obtained results, while secondary results are detailed in the Appendix A .

3.3.1 Gaze behaviors

According to our objective and hypotheses, we focused on five comparisons, related to three categories: (1) the stare-in-the-crowd effect in static conditions, (2) *catching someone else staring* and (3) *being caught staring* phenomena, in line with Crehan et al. [2015].

— For (1), we compared the averted to the directed gaze conditions – A vs. D. Then, we compared each static condition with each dynamic.
— For (2), averted versus averted-then-directed – A vs. AD, and directed versus averted-then-directed – D vs. AD.
— For (3), the averted versus directed-then-averted – A vs. DA, and directed versus directed-then-averted – D vs. DA.

For pairwise comparisons, we ran dependent t-tests for paired samples on the six metrics we described in Section 3.2.6 as continuous variables. Such tests guarantee conservative results in the comparison between different gaze conditions. The normal distribution as-

sumption was verified for 25 of our 30 dependent paired samples when running a Shapiro-Wilk test: we ran Student's t-tests for these samples, and Wilcoxon signed rank tests for the remaining ones. Due to our multiple comparison design, we conducted a Bonferroni correction which changed our target significance level from α=0.05 to α=0.00166. We detail here two representative results out of the five comparison see Figure 3.5: one static A vs. D, and one dynamic case A vs. AD (the rest of the detailed results are presented in the Appendix A).

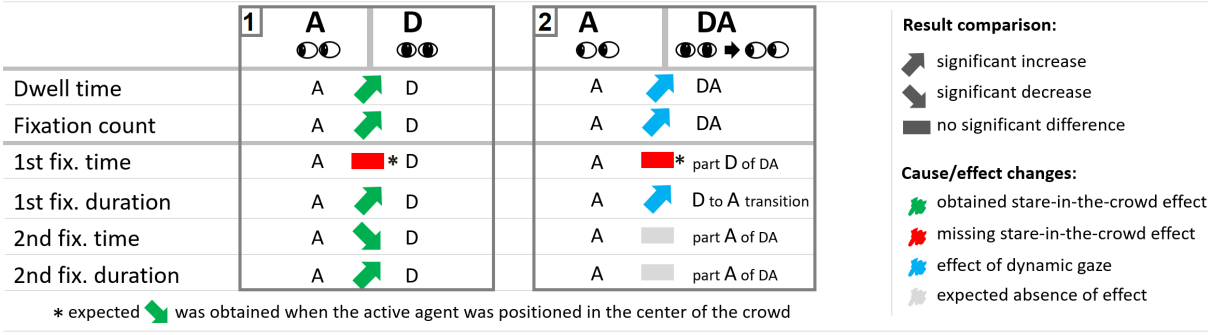


Figure 3.5 – Summary of results for two representative cases of comparison: 1) A vs. D reveals the presence of the stare-in-the-crowd effect in virtual reality, and 2) A vs. DA reveals effects of dynamic gazes.

For each metric, results present the means and standard deviations, along with significance level, plus statistics and effect size (both when doing Student's t-test). Results are shown by comparison of pairs and they are based on the averages obtained by each user *across all trials that share the same gazing conditions regardless of position*, *i.e.*, 18 in total for each condition. In these tables, a symbol * indicates a p-value <0.00166, ** a p-value <0.00033, and *** a p-value <0.00003.

Comparison A vs. D: interpretation. As shown in Table 3.1, p-values from the metrics *dwell time, fixation count, first and second fixation durations* were all significant, with higher values for the directed condition, which are all **indicators of the presence of a stare-in-the-crowd effect**. We also expected users to spot the active agent in the directed gaze condition sooner, which should be reflected through significantly earlier first fixation time. Such results have been reported and used to confirm the presence of a stare-in-the-crowd effect in previous studies with drawing or photographic stimuli [Ramamoorthy et al., 2019; Von Grünau and Anston, 1995]. In our experiment, *first fixation time* results do not reveal such a significant difference. However, when computing the

Table 3.1 – Gaze metrics results - comparison of A vs. D conditions

	Averted		Directed			
Metric	Mean (SD)		Mean (SD)	p-value	t	η_p^2
Dwell time	504 (175)	↗	1570 (864)	<0.00001 ***	-6.75	0.61
Fixation count	1.15 (0.29)	↗	2.35 (1.03)	<0.00001 ***	-6.45	0.59
1st fix. duration	332 (77)	↗	552 (185)	<0.00001 ***	-5.61	0.52
1st fix. time	5173 (1213)		4969 (1402)	0.53119	0.63	0.014
2nd fix. duration	407 (282)	↗	554 (214)	0.00158 *	wilc.	wilc.
2nd fix. time	8602 (1395)	↘	6861 (1785)	0.00031 **	4.09	0.37

Time and duration in ms.

analysis based on the nine positions of the active agents, we find a significant difference in the *first fixation time* metric for the three central positions in the crowd – in **accordance with the presence of a stare-in-the-crowd effect** – while no significant difference on the sides. We believe that this phenomena is caused by the visual difference of the proposed apparatus, virtual reality, in our case, and photographic images on a screen in previous studies. Indeed, virtual reality allows for a wider field of view (100° on the FOVE), that we filled up for better immersion, compared with the simple image (that is limited to approximately 30°) and that can justify why peripheral stimuli were detected slower. Further details on this analysis are provided in the paragraph **Active agent's position effect** of Appendix A. Based on the expectations of the stare-in-the-crowd effect, our results nonetheless show a significantly earlier *second fixation time* on the directed condition compared to the averted one, following the trend expected for the first fixation time. Figure 3.5 (left) summarises the comparison between the results on averted and directed conditions and its interpretation for the stare-in-the-crowd effect.

When comparing with Crehan et al. [2015], we found the same results on all our metrics, except for the significantly longer duration for the first fixation in the directed condition in our experiment. Nonetheless, this result is in line with other previous studies [Ramamoorthy et al., 2019; Von Grünau and Anston, 1995] and the stare-in-the-crowd effect by definition. In addition, it could be explained by a stronger effect of virtual reality to capture attention with directed gazes, as suggested by our larger effect size results for the other metrics, compared to Crehan et al.'s ones [2015].

Table 3.2 – Gaze metrics results - comparison of A vs. DA conditions

	Averted		Directed-then-Averted			
Metric	Mean (SD)		Mean (SD)	p-value	t	η_p^2
Dwell time	504 (175)	↗	808 (363)	0.00005 **	-4.78	0.44
Fixation count	1.15 (0.29)	↗	1.56 (0.56)	0.00015 **	-4.37	0.40
1st fix. duration	332 (77)	↗	483 (165)	0.00003 ***	-4.92	0.45
1st fix. time	5173 (1213)		4847 (1307)	0.32902	0.99	0.03
2nd fix. duration	407 (282)		374 (95)	0.34921	wilc.	wilc.
2nd fix. time	8602 (1395)		7773 (1724)	0.05175	2.03	0.12

Time and duration in ms.

Comparison A vs. DA: interpretation. As shown in Table 3.2, *first fixation duration, dwell time and fixation count* were significantly different between averted and directed-then-averted conditions, with higher values in the latter. In contrast, *second fixation duration* and *second fixation time* were not significantly different between these conditions. *First fixation time* metric did not show significant differences either, but, as for the previous case, significant differences are present for central positions, for further details see **Active agent's position effect** paragraph in Appendix A. The results for all the other five metrics might be understood and explained according to the procedure of the directed-then-averted gaze trial. Indeed, in this condition, once the first fixation had started on the active agent, users could observe a dynamic gaze change. This might have captured their attention and could explain the fact that they stared significantly longer towards the active agent during the first fixation. After, the active agent entered the averted gaze condition: this could explain why the directed-then-averted condition results of second fixation duration and second fixation time were not significantly different compared to the averted condition ones. Finally, dwell time and fixation count were nevertheless significantly higher in the dynamic condition, which could be explained by the multiple rechecks by users towards the active agent during the remaining time of a trial, to see if the agent would look at them again. Figure 3.5 (right) summarises the comparison between the results on the averted and directed-then-averted conditions and its interpretation in relation with the stare-in-the-crowd effect and the effect of dynamic gaze changes.

In addition, when comparing our results to the ones of Crehan et al. [2015], both studies found similar effects, except that in their case instead of finding a significant

difference for the first fixation duration, they found it for the second fixation one.

3.3.2 Gaze behaviors and social anxiety

To investigate whether users with a higher level of social anxiety were less likely to gaze towards agents who are gazing at them, we computed correlations between the final score on the social anxiety questionnaire, *i.e.*, the Liebowitz Social Anxiety Scale, and our gaze metric data. This final social anxiety score ranges from 0 to 144, with a low score depicting an absence of social anxiety and high score depicting a significant presence of social anxiety. As some of our variables were not normally distributed (Shapiro-Wilk test), we conducted Spearman's rank-order correlation on our data, between the final social anxiety scores and the gaze metrics results to be able to compare the correlation coefficients between themselves.

As expected, we found some negative correlations between social anxiety and metrics of the eye-tracking data. In particular, dwell time for directed (D) and dynamic conditions (DA, AD) showed significant negative correlations ($D : r_s = -0.42, p = 0.022, AD : r_s = -0.57, p = 0.001, DA : r_s = -0.37, p = 0.047$), indicating that the more socially anxious the user was, less time he or she spent observing the agent whose gaze was directed towards them. The correlation was particularly high in the AD condition (*getting caught staring*). Other metrics were not correlated with social anxiety, except for the averted condition first fixation duration ($A : r_s = -0.40, p = 0.028$) and the averted-then-directed condition fixation count ($AD : r_s = -0.49, p = 0.006$).

3.4 General discussion

This study evaluated virtual reality users' gaze behaviors depending on different gaze conditions that were applied to a virtual crowd, and therefore aimed to test the stare-in-the-crowd effect in virtual reality. Our H1 hypothesis was that the stare-in-the-crowd effect would be preserved in virtual reality, and H2 hypothesis that we would observe a negative correlation between the time spent towards the agents who are staring at the user and the user's level of social anxiety.

Hypoteses. In terms of verifying H1, we compared our results with the one obtained by Crehan et al. [2015] using similar metrics, and found similar effects, confirming the stare-in-the-crowd effect in virtual reality. Some differences with the previous study were

also found, which were discussed in section 3.3.1. One major difference was that we used a virtual reality environment that could have affected the gaze behavior simply due to the field of view being different than in studies using photographs. It appears to be important how the user is positioned in virtual reality as well, since some aspects of the stare-in-the-crowd effect were not present for characters outside the central region. An important difference was also between dynamic conditions of both studies. In our study, we found less gaze fixations in the dynamic conditions than in the directed static one, oppositely to the findings of the previous study. This could be explained by users expecting changes in the behavior of virtual agents in virtual reality, since agents were slightly animated (blinking), whereas photographic stimuli may not have had the same anticipation effect. We believe that our results are potentially more accurately transferable to physical reality than previous results that were collected by using photographs only.

Regarding H2, our results show that social anxiety is negatively correlated with dwell time for all conditions that include directed gaze. Therefore, on average, the higher the social anxiety, the less time users spent looking at the agents when their gaze was directed towards them, which is in line with the gaze behavior of socially anxious individuals [Baker and Edelmann, 2002]. Particularly interesting is the result that the averted-then-directed condition ("being caught staring") had the strongest correlation compared to other conditions, meaning that socially anxious individuals were particularly sensitive to agents who looked at them after they saw them. Other metrics (fixation time, etc.) were not correlated, meaning that perhaps the additive effect of dwell time metric was stronger. However, we did get a negative correlation with fixation count for the averted-then-directed condition again, but also for the averted condition, with the first fixation duration. The latter could indicate that users with higher social anxiety may avoid to look at characters at the very beginning of the trial for fear of meeting their gaze. Some users reported their fear of the virtual agents in our post-experiment questionnaire and reported avoiding agents who were staring at them: *"actually, older people are super scary", "embarrassed by the stare of the avatars towards me, I run away from them rather quickly", "some avatars felt creepier than others, their gaze felt heavier when they were looking for afar, and more normal or natural when they were actually just in front of me".* Importantly, we were able to demonstrate a stare-in-the-crowd effect in our study, indicating that the amount of socially anxious individuals in our sample of users was not high.

Limitations. Firstly, our sample of participants was not balanced in terms of gender, which may have affected our data. However, we made sure that we had a balanced representation of both genders in the stimuli sample (virtual characters). We also cannot generalise our results to more natural social situations. While we designed the agents to be as realistic in appearance as possible, the integration of better models and animations could be used to make the results more transferable to interactions in the physical world. In addition, other scenarios than the one where the virtual audience is listening to a speaker could be considered. Moreover, in this study we took behavioral measures using an eye-tracking system and an indirect measure with the social anxiety questionnaire, however we could also have used some subjective measures such as presence and social presence [Bailenson et al., 2001; Slater et al., 1994]. Another limitation is that we did not check specifically for cybersickness. Nonetheless we ensured a sufficient framerate in the FOVE headset and our virtual reality users were seated and had limited movements, therefore adverse effects of cybersickness were limited. We also found the importance of where the user is positioned in virtual reality as this affects the stare-in-the-crowd effect. Future studies are needed to better understand the stare-in-the-crowd effect at different observing positions and also in times when the user is allowed to move through the environment.

3.5 Conclusions and future work

With this study we demonstrate the presence in virtual reality of the well-known stare-in-the-crowd effect, which predicates the existence of a search asymmetry between directed and averted gaze towards the observer, with faster detection and longer fixation towards directed gaze. In other words, it represents the tendency of humans in noticing and observing, more frequently and for longer time, gazes oriented toward them (directed gaze) than gazes directed elsewhere (averted gaze). We also demonstrate that this effect is milder with people reporting higher social anxiety levels.

With this, we showed that gaze can indeed change the focus of attention of a user, and potentially trigger the interaction with an agent. Such promising results help the understanding of social interactions in virtual reality applications and the design of more engaging experiences with virtual agents. For example, our gaze conditions could be used to initiate the interaction with the user in a virtual crowd while, in the previous chapter, we demonstrated how body motion editing can help improve the understanding of user

intentions during the interaction. We also demonstrated a simple dynamic gaze condition that signals complex social behavior, *e.g.*, directed-then-averted gaze could potentially be interpreted as a sign of embarrassment of the agent. These subtle gaze conditions could be explored further to create more believable social interactions in virtual reality.

In the future, we plan to explore the stare-in-the-crowd and other related effects in more complex scenarios, *e.g.*, including more dynamic and heterogeneous virtual agents, changing their number, giving the user different tasks. Moreover, we will expand our analysis to also consider further social and behavioral aspects of our human users, so as to see how they relate to the gazing times.

Chapter 4

Haptic Rendering of collision in a virtual crowd

Contents

To continue the exploration of interaction between human and virtual humans, we present another user study, on which I collaborated not as primary investigator. My contribution is limited to development of some functionality, collaboration to the investigation and the writing.

As we saw, Virtual reality is a valuable experimental tool for studying human behavior. For this reason it is also used to study human movement and mutual displacement. For example, in crowd modelling and simulation for analysing the dynamics of local interactions between individuals in crowded environments [Bruneau et al., 2015]. However, this is not an easy task, since human movement, and in general human behavior, relies on many different variables and covers a wide range of interactions. Indeed, there is a

Figure 4.1 – Our objective is to understand whether and to what extent providing haptic rendering of collisions during navigation through a virtual crowd (right) makes users behave more realistically. Whenever a collision occurs (center), armbands worn on the arms locally vibrate to render this contact (left). We carried out an experiment with 23 participants, testing both subjective and objective metrics regarding the users' path planning, body motion, kinetic energy, presence, and embodiment.

high chance that participants will behave differently from how they behave in real life. This mismatch is due to different reasons, most of them linked to the limited sense of realism that virtual reality provides. In this respect, significant efforts have been put in the evaluation of the realism of behaviors in virtual reality [Cirio et al., 2013; Olivier et al., 2017], in the understanding of the visual information required [Lynch et al., 2017] or in the development of highly-realistic virtual environments and characters [Achenbach et al., 2017]. Still most virtual reality experiences lack any haptic sensation, which is of course of paramount importance when studying crowd behavior and interactions. For example, if we are unable to render the sensation of bumping into virtual characters when navigating in a crowded environment, participants might stop avoiding collisions, leading to data that does not capture well how humans truly behave. For this reason, studies of collective behavior in virtual reality are often limited to cases considering distant interactions only [Rio and Warren, 2014; Rio et al., 2018], so as not to require any haptic feedback.

This study explores the role of contact interactions (collisions) during navigation in a crowded environment (see Figure 4.1). To do so, we employ a set of wearable haptic interfaces able to provide compelling vibrotactile sensations of contact to the user's arms. Our objective is to investigate whether and to what extent the rendering of contacts influences the user's behavior in this context, as well as limits the occurrence

of certain well-known artifacts, such as when the user's virtual avatar interpenetrates other virtual characters. We conducted an experiment (N=23) where participants were equipped with four wearable haptic interfaces (two on each arm), and asked to navigate in a densely-crowded virtual train station. We evaluate objective metrics related to the user's behavior with respect to the crowd, as well as subjective metrics related to the user's sense of presence and embodiment. First, we carried out the experiment without haptic rendering of contacts, then with haptic rendering, and finally once again without haptic rendering. This experimental design enables us to register the difference in user's behavior when activating the haptic feedback as well as the persistence of any relevant after-effect. These results are expected to help all researchers planning to use virtual reality for studying human behavior when navigating a crowded environment. My contribution in functionalities is related to the development and study of collision volumes of interpenetration, that we will detail in this chapter.

The rest of the chapter is structured as follow. Section 4.1 describes the experimental setup, methods, and task. Section 4.2 presents the metrics we considered, based on the study of local body movements, trajectories, energy, contacts, embodiment, and presence. Section 4.3 discusses the results and analyses the differences between the considered conditions using inferential statistical analysis methods. Section 4.4 discusses our findings as well as their implications for crowd experiments in virtual reality. Finally, Section 4.5 draws the final remarks and discusses future work on the topic.

4.1 Experimental overview

The purpose of this study is to investigate the effect of haptic rendering of collisions on participants' behavior during navigation through a static crowd in virtual reality. To explore this question, we immersed participants in a virtual train station and asked them to perform a navigation task which involved moving through a crowd of virtual characters. In some conditions, collisions with the virtual characters were rendered to participants using 4 wearable vibrotactile haptic devices (actuated armbands). Our general hypothesis is that haptic rendering changes the participants' behavior by giving them feedback about the virtual collisions. Moreover, we also expect that even after removing haptic rendering, an after-effect still persists on the participants' behavior.

4.1.1 Materials & methods

Apparatus

For the purpose of immersing participants in the virtual environment and investigating the potential effects of haptic rendering while navigating in groups of characters, we used the following devices, which are summarized in Figure 4.2:

— **Motion Capture**: to record participants' body motions, as well as to render their animated avatar in the scene, we used an IMU-based (Inertial Measurement Unit) motion capture system (Xsens [1]).

— **HMD**: to immerse participants in the virtual environment, we chose to use a Pimax [2] virtual reality headset, in particular because of the wide field of view provided in these situations of close proximity with other characters (specifications: 90 Hz, $200° fov$, 2560 × 1440 resolution). The HMD was used with 4 SteamVR 2.0 base stations, providing a tracking area of approximately 10×10 m. This setup enabled participants to physically walk in the real space, while their walking movements were displayed on their avatar in the size-matched virtual environment.

— **Haptic Rendering**: to render haptic collisions between participants and the virtual characters, we equipped participants with four armbands (one on each arm and forearm) [Scheggi et al., 2016]. Each armband is composed of four vibrotactile motors with vibration frequency range between 80 and 280 Hz and controlled independently. Motors are positioned evenly onto an elastic fabric strap (see Figure 4.2 in red). An electronics board controls the hardware. It comprises a 3.3 V Arduino Mini Pro, a 3.7 V Li-on battery, and a Bluetooth 2.1 antenna for wireless communication with the external control station.

— **Computer**: to let participants move freely in the environment, they were equipped with a MSI VR One backpack computer, which was running the experiment. All the devices were connected directly to this computer (specifications: NVidia GTX 1070, Intel Core i7-7820HK processor, 32 GB RAM).

Haptic rendering

Haptic rendering requires collisions to be detected in the virtual environment. Since haptic devices were worn on participant's arms, we detected collisions between their avatar

1. https://www.xsens.com/
2. https://www.pimax.com/

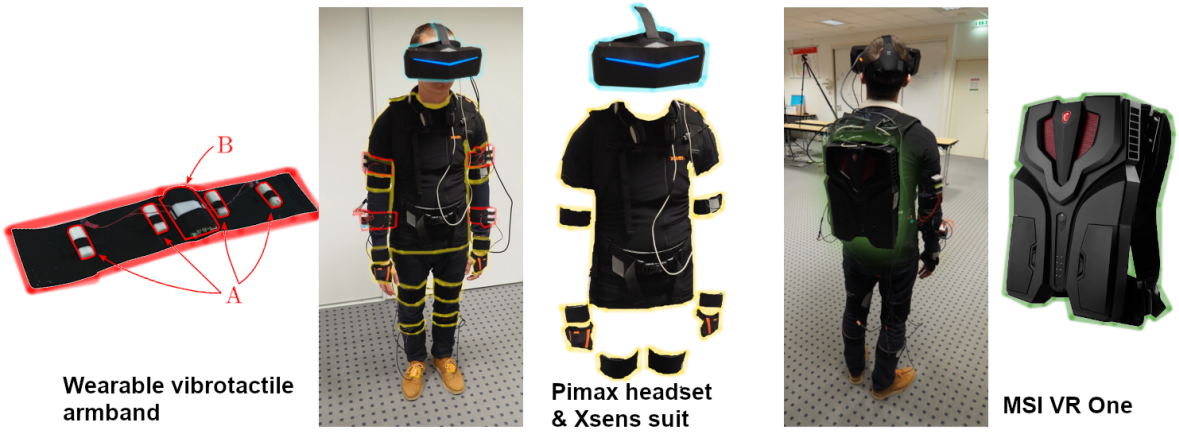


Figure 4.2 – Devices worn by participants during the experiment. In red is highlighted the Wearable vibrotactile armband, composed of four vibrating motors (A); the electronics is enclosed in a 3D-printed case (B) [2016]. In yellow we highlight the Xsens suit, in blue the Pimax headset and in green the backpack computer, MSI VR One.

(animated using the Xsens motion capture system) and the virtual crowd. To this end, we segmented each avatar's arm into three parts (arm, forearm, and hand), and attached to each segment a Unity capsule collider that reported on collisions with other objects in the scene (see Figure 4.3). When a collision was detected, that is if one of the six segments of the avatar entered in collision with the geometry of any virtual crowd character, one of the four haptic devices was activated. More specifically, colliders on the left (resp. right) virtual forearm and hand activated the armband located on participants' left (resp. right) forearm, while colliders on the left (resp. right) virtual upper arm activated the armband located on participants' left (resp. right) upper arm.

In terms of vibrations, each vibro-motor of an armband was driven using a single parameter called *vibrotactile rate*, which controlled both the amplitude and the frequency of vibration. During the experiment, all the motors of an activated armband were therefore controlled using the same *vibrotactile rate*, which varied according to a 10 Hz-period sine wave profile. The variation of the vibrotactile rate resulted in a frequency of vibration in the range of [57–126] Hz. Although these motors can vibrate up to 255 Hz, we decided to limit their range after participants in a pilot study reported the full vibrating range to be too strong.

Communication with the armbands was performed at 4 Hz, meaning that collisions with a duration less than 250 ms were not rendered to participants, and that there was a maximum delay of 250 ms in activating (resp. stopping) the armbands after a collision

was detected (resp. ended).

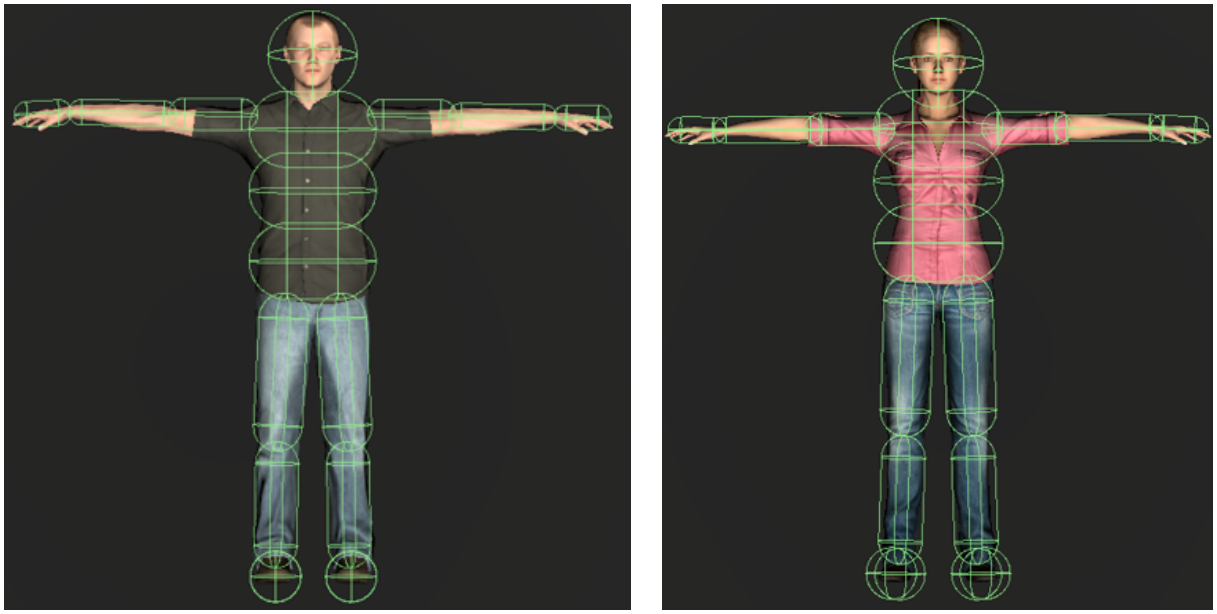
4.1.2 Environment & task

Participants were immersed in a digital reproduction of the metro station "Mayakovskaya" in Moscow, amongst a virtual static crowd (see Figure 4.4). A total of 8 different configurations of the scene were prepared in advance and used in the experiment. A configuration is defined by the exact position of each crowd character in the virtual station. In each configuration, the crowd formed a squared shape, and character positions followed a Poisson distribution resulting in a density of 1.47 ± 0.06 character/m^2. Such a distribution combined with such a level of density ensures that a gap of 0.60 m on average exists between each character. The crowd is composed of standing virtual characters animated with various idle animations (only small movement but standing in place). In each configuration, characters were animated according to two types of behavior, either waiting (oriented to face the board displaying train schedules, moving slightly the upper body) or phone-calling (with a random orientation). We used several animation clips for each of the two behaviors, in order to prevent the exact same animation clip to be used for two different virtual characters.

At the beginning of each trial, participants were initially standing at one corner of the square crowd, embodied in a gender-matched avatar (see Figure 4.3). They were instructed to traverse the crowd so as to reach the board displaying train schedules, and to read aloud the track number of the next train displayed on the board before coming back to their initial position. They were physically walking in the real room, while their position and movements were used to animate their avatar. This task required participants to reach the opposite corner of the space in order to read information on the board, while forcing them to move through the virtual crowd. Also, the screen displayed the train information only when participants were at less than 2 m from it (i.e., when they reached the green area displayed in Figure 4.4.b). Furthermore, we provided the following instruction to participants prior to the experiment: *"Walk through the virtual train station as if you were walking in a real train station"*.

4.1.3 Protocol

Upon arrival, participants were asked to fill in a consent form, during which they were presented the task to perform. They were then equipped with the equipment listed in



a) b)

Figure 4.3 – Male (a) and female (b) avatars used to represent the participants in the virtual environment. For both avatars the capsule around each segment represents the solid used to compute collisions.

Section 4.1.1. Calibration of the Xsens motion capture system was then performed to ensure motion capture quality, as well as to resize the avatar to participants dimensions. Once ready, participants performed a training trial in which they could explore the virtual environment and get familiar with the task.

The experiment then consisted of 3 blocks of 8 trials, where the blocks were presented for all participants in the following order: *NoHaptic1*, *Haptic*, and *NoHaptic2*. The *Haptic* block corresponded to performing the task with haptic rendering of contacts, while the *NoHaptic* blocks did not involve any haptic rendering of contacts. The experiment therefore consisted in performing first a block without haptic rendering, in order to measure a baseline of participants' reactions. The purpose of the second block was then to investigate whether introducing haptic rendering influenced their behavior while navigating in a crowd, while the purpose of the last block (without haptic) was to measure potential after-effects. In each trial, participants performed the task described in Section 4.1.2 once. Each block was comprised of 8 trials, corresponding to the 8 crowd configurations presented in Section 4.1.2, performed in a random order. At the end of each block, participants were asked to answer the *Embodiment* and *Presence* questionnaires (Tables B.2,

a) b)

Figure 4.4 – Snapshots of the environment under two different points of view. Participants started from the blue cross on the floor, and were instructed to reach the screen board. Figure (b) displays an example trajectory in a red doted line. The screen displayed the train information only when participants reached the green area.

B.3, B.4 & B.5) while remaining in the virtual environment. Finally, at the end of the experiment, participants filled in a demographic questionnaire.

4.1.4 Participants

Twenty-three unpaid participants, recruited via internal mailing lists amongst students and staff, volunteered for the experiment (8F, 15M; age: avg=26 ± 6, min=18, max=43). They were all naive to the purpose of the experiment, had normal or corrected-to-normal vision, and gave written and informed consent. The study conformed to the declaration of Helsinki, and was approved by the Inria internal ethical committee (COERLE).

4.1.5 Hypotheses

We proposed a set of hypotheses to evaluate how participant behaviors would change with haptic rendering.

*H*1: Haptic rendering will not change the path followed by participants through the crowd. Indeed, pedestrians mainly rely on vision to control their locomotion [Patla, 1997; Warren, 1998], and we replicated each crowd configuration across the 3 blocks, resulting in identical visual information for participants to navigate. Therefore, the followed path will be similar in the tree blocks of the experiment

(*NoHaptic1*, *Haptic* and *NoHaptic2*).

$H2$: Haptic rendering of collisions will make participants aware of collisions and influence their body motion during the navigation through the crowd. Therefore, concerning the *NoHaptic1* and *Haptic* blocks of the experiment, we expect that:

$H2_1$: Participants will navigate in the crowd more carefully in the *Haptic* block in order to avoid collisions. There will be more local avoidance movements (e.g., increased shoulder rotations) and a difference in participants' speed.

$H2_2$: With theses changes on participants' local body motions, there will be both less collisions, and smaller volumes of interpenetration when a collision occurs.

$H3$: We expect some after-effect due to haptic rendering, i.e., we expect that participants will remain more aware and careful about collisions even after we disabled haptic rendering. Therefore we expect $H2_1$ and $H2_2$ to remain true in the *NoHaptic2* block.

$H4$: Haptic rendering will improve the sense of presence and the sense of embodiment of participants in virtual reality, as they will become more aware of their virtual body dimensions in space with respect to neighbour virtual characters.

4.2 Analysis

During the experiment, we recorded at 45 Hz the trajectories of participants, as well as the position and orientation of their limbs in the virtual environment using the Xsens sensors and Unity. We also recorded the body poses over time of each character of the virtual crowd. Then, we were able to replay offline the entire trials in order to compute complex operations such as the volume of each collision.

4.2.1 Metrics

We use this data to analyze various variables to validate our hypotheses. We detail here only the data relative to collisions, on which I had a major role as contributor. The other metrics are then briefly presented, and further detailed in Appendix B.

Collisions. A collision is the *detected contact* between any part of the participant's virtual body and any part of the mesh of one virtual character. We identify a collision by the pair participant-virtual character as well as the initial time. This means that we separately classify collisions with different characters, even if they are happening at the

same time. This also means that we can detect several collisions with the same character but with different initial times. The detection starts at the first contact of any of the limbs of the character involved and the participant's geometry, and it lasts until there is no more contact detected between the two respective meshes. The whole collision computation scheme is summarized in Figure 4.5. To analyze the collisions we selected two main values of interest: the *number of collisions* and the *maximum volume of interpenetration* between the participant and the virtual character during a collision:

— *Number of collisions.* We count any collision with an interpenetration volume greater than 10^{-6} m^3 and lasting more than 10 ms.

— *Maximum volume of interpenetration.* The maximum volume of interpenetration between a participant's avatar and a virtual character during a collision is computed at each time stamp through the voxelization of the intersection of their respective meshes, according to the following procedure. Each 10 ms the computation starts from the meshes of the two characters involved. Around those, we build an AABB (axis aligned bounding box), which is then iteratively subdivided in octant where, at each of this octant-iteration, only the voxels in collision are kept. The octant-iteration stops when the target voxel size is reached. In our analysis, it was set to a cube of width 0.01 m. This process is shown in Figure 4.6. At the end we collect all the volumes computed at each time interval of 10 ms and we extract the maximum one.

Others. In addition to the collisions and in order to study $H1$, we have compared trajectories, representing them as sequences of traversed cells, which where defined on the environment based on Delaunay triangulation [Chew, 1989] between agents of the crowd. Furthermore, we have studied the body motions (to explore $H2_1$) decomposed in (i) shoulder rotations, computed as the shoulder orientation while passing through two close agents, and (ii) walking speed. Finally, we evaluated Presence and Embodiment with two questionnaires that helped us evaluate hypothesis $H4$.

4.2.2 Statistical analyses

Our objective is to understand whether and to what extent users change their behavior in each experimental block. To do so, we analyzed the differences across blocks for all the aforementioned variables. For all dependent variables, we set the level of significance to $\alpha = 0.05$. First, a Shapiro-Wilk test was performed to evaluate whether the distribution

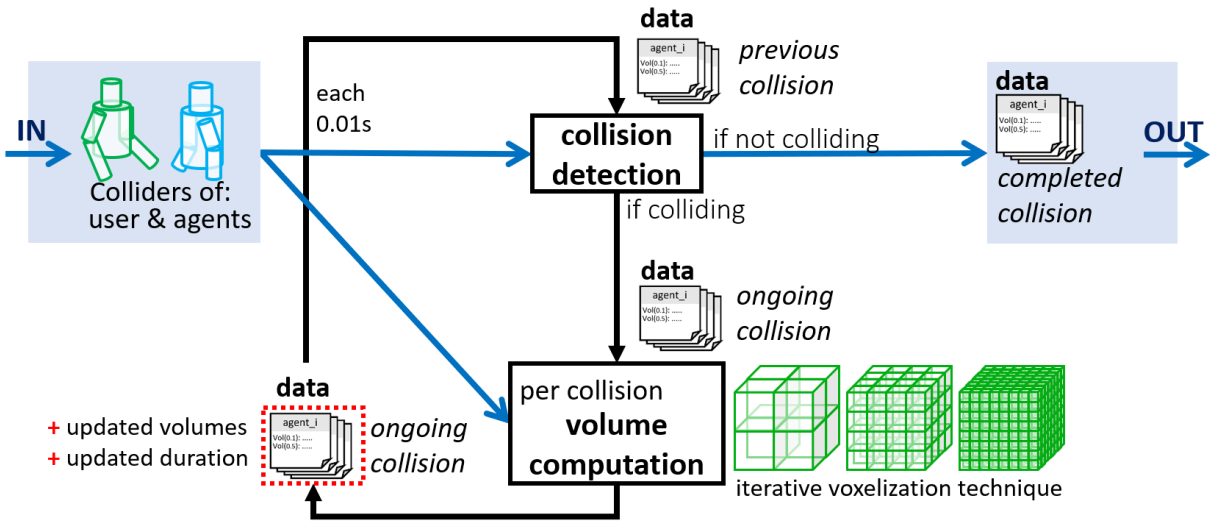


Figure 4.5 – Collision iteration loop scheme representing one step of the collision detection, in which we detect if there is a collision (either a new or an ongoing one) and compute its volume. We add this information to collision's data. When the collision is finished we send out the data.

of our data followed a normal distribution. If the distribution was not normal, a Friedman test was performed to evaluate the effect of the condition on these variables. Post-hoc comparisons were then performed using a Wilcoxon signed rank test with Bonferroni correction. On the other hand, if the distribution was normal, a one-way analysis of variance (ANOVA) with repeated measures was performed. Greenhouse-Geisser adjustments to the degrees of freedom were applied if the data violated the sphericity assumption. Bonferroni post-hoc tests were used to analyze any significant effects between groups.

4.3 Results

As for the previous section, we will only detail here the results related to the collision values. For more details on the remaining results please refer to Section B.2 of the Appendix.

Collisions. Figure 4.7 illustrates the results regarding collision characteristics, i.e., number of collisions as well as volume of interpenetration. The average number of collisions per trial was influenced by haptic rendering with a large effect ($F(2, 44) = 7.13, p = 0.002, \eta_p^2 = 0.25$). Post-hoc analysis showed that the number of collisions was higher dur-

a) b) c)

Figure 4.6 – Volume computation using iteration of voxel spaces of decreasing dimensions. (a) Starting from the AABB (axis aligned bounding box) around the selected geometries, the first voxel space with 8 voxels (green cubes) is created and intersected with the geometries. (b) In the next iteration only the intersecting voxels are kept, and further subdivided into 8 cubes each. (c) The process is iteratively applied until reaching the minimum subdivision size, where the final interpenetration volume is displayed in purple.

ing the *NoHaptic1* block (71 ± 29.2) than during the *Haptic* (62.8 ± 34.6, $p = 0.018$) and *NoHaptic2* blocks (60.7 ± 34.6, $p = 0.002$), which shows that participants made on average more collisions before they experienced haptic rendering. The average volume of interpenetration was also influenced by the block ($F(2, 44) = 4.35, p = 0.019, \eta_p^2 = 0.16$), where post-hoc analysis showed that this volume was smaller ($p = 0.016$) in the *Haptic* block (0.6 ± 0.3 dm^{-3}) than during the *NoHaptic1* (0.8 ± 0.3 dm^{-3}).

These results validate our hypothesis $H2_2$, that states that haptic rendering reduces the severity of collisions between participants and virtual characters. Furthermore, as the number of collisions is higher during block *NoHaptic1* than during block *NoHaptic2*, this also supports $H3$ on potential after-effects of haptic rendering.

Other Results. Regarding the other results, we verified $H1$ (haptic rendering will not affect the path followed by the participant) with the studies of trajectories, as well as identified a significant difference in shoulder rotation, supporting $H2_1$, as, participants tended to rotate more to squeeze in narrow passage between the crowd. We also observed an after-effect with a significant difference in block *NoHaptic2* related to *Haptic* with less collisions and more shoulder rotations, supporting $H3$. Additionally, in support of $H2_1$, we showed that haptic rendering has an effect on walking speed, which was lower in the presence of haptic feedback, compared to higher speed in block *NoHaptic1* and

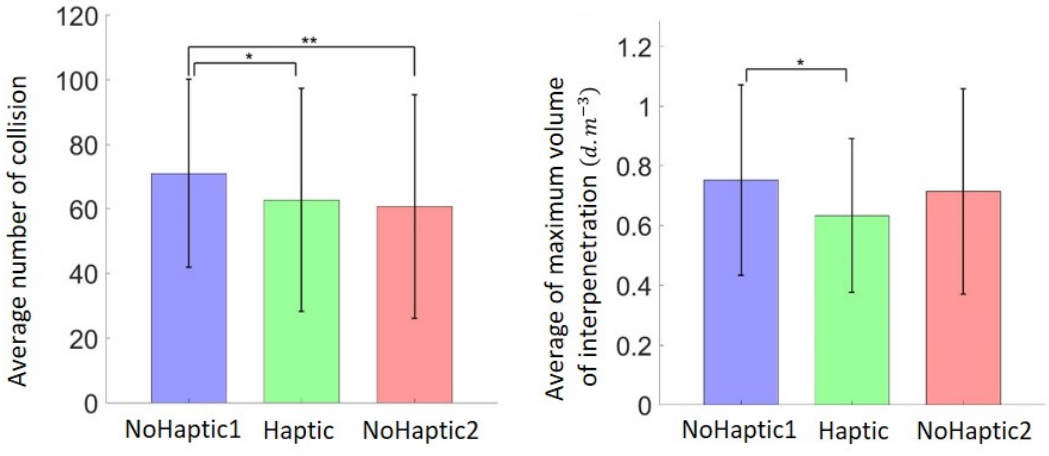


Figure 4.7 – Main significant differences between the three blocks of the experiment (*NoHaptic1*, *Haptic* and *NoHaptic2*) number of collisions per trial(left) and volume of interpenetration (right). Error bars depict standard deviation of the mean.

NoHaptic2. On the other hand, results for the questions of presence and embodiment did not present any significant difference between the experiment's blocks, not supporting $H4$. All these results are described in details in Appendix B.

4.4 Discussion

The main objective of this study was to evaluate the effect of haptic rendering of collisions on participants' behavior while navigating in a dense virtual crowd. To this end, we designed an experiment where participants had to reach a goal by physically walking in a virtual train station populated with a dense crowd. Participants were equipped with vibrotactile sensors located on their arms and performed this task following 3 blocks: *NoHaptic1*, *Haptic* and *NoHaptic2*, for which haptic rendering of collisions with virtual characters was not experienced, experienced, and not experienced again, respectively.

We discuss here the results related to collisions and an overview of the others, further discussions can be found in Section B.3 of the Appendix. Being more cautious effectively resulted into less collisions as expected in hypothesis $H2_2$. Results presented in Section 4.3, paragraph Collisions, show that the average number of collisions as well as the average volume of interpenetration were significantly lower in the *Haptic* block than in the *NoHaptic1* block. Furthermore, this observation is consistent with previous studies [Louison et al., 2018] where haptic feedback lowered the number of collisions with a static object.

Regarding the other results, trajectory data are aligned with the idea that vision is the most prominent stimuli human used to define their path [Patla, 1997; Warren, 1998], supporting H_1. In this experiment, we also demonstrated that haptic rendering had an effect on shoulder rotations, which supports hypothesis $H2_1$. In particular, participants rotated more their shoulders when traversing the gaps between virtual characters during the *Haptic* block than during the *NoHaptic1* block. This result is consistent with the observations of Mestre et al. [Mestre et al., 2016] with participants passing through a virtual half-open door with or without haptic rendering. More generally, let us recall that the human trunk is most often larger along the transverse axis than along the antero-posterior axis. Thus, the more the participants turn their shoulders the smaller the volume swept by their body motion. Regarding $H4$, we have not found any significant effect of haptic rendering on Embodiment and Presence. We think this can be caused by the inaccurate way in which we simulate contact (vibrotactile feedback vs. contact sensation) or the precision of the contact point, for additional details check Section B.3 in the Appendix.

Haptic rendering after-effects. While there were less collisions and more shoulder rotations observed in the *Haptic* block in comparison with the *NoHaptic1* block, there was no difference between the *Haptic* and the *NoHaptic2* blocks. This supports hypothesis $H3$ on potential after-effects of haptic rendering. However, such an after-effect did not equally influence all measurements, such as walking speed that increased again in the *NoHaptic2* block. One possible explanation might be a perceptual calibration of the participants. During the experiment, participants became more familiar with the environment, the task to be performed, but also the virtual representation of their body and the virtual environment, enabling them to move faster and better avoid collisions with the virtual characters in the last block (*NoHaptic2*).

Another point to highlight is that participants, at the beginning of the *Haptic* block, did not know that contacts would now trigger a vibrotactile haptic sensation. For this reason, we might expect to see a short learning phase at the beginning of the block, where participants learn to deal with the newly-rendered haptic collisions. Considering this point, we can expect the effect of providing haptic sensations of collisions even stronger than registered. However, to provide a more definitive conclusion on the role of the haptic after-effect would require to add a control group with no haptic rendering throughout the 3 blocks of the experiment, which could be explored in future work.

These results can also open perspectives regarding the design of new experiments in-

cluding haptic priming tasks. In a recent study, Krum et al. [2018] showed that haptic priming of collision had no effect on participants' proxemics and more precisely on distances with a virtual character. It is important to note that the task was different: it included an interaction with one virtual character and there was no risk of collision since the virtual character never came very close to the participant. It would be interesting then to re-evaluate such influence when the intimate space is violated by a virtual character.

Limitations. Our study had a few limitations. For example, we employed a limited number of haptic rendering devices located on participants arms only. It is quite possible that employing more devices, including some for the legs and hips, would have resulted in stronger effects. However, our setup still revealed significant effects, and the question of nature, number, and location of haptic devices would probably require a fully dedicated study. Another related issue is the *quality* of the provided haptic sensations. Our devices show high wearability and portability, but can only provide vibrotactile haptic sensations. Other haptic delivery options include the use of arm or full-body exoskeletons, which can provide well-rounded force sensations. However, these devices are significantly more cumbersome and expensive than those employed in this work, severely limiting their applicability and availability.

A second limitation concerns the behavior of the virtual characters present in the crowd. Indeed, they do not react to collisions, as noticed by some participants in their feedback. It would therefore be required to have an animation technique capable of reacting to collisions such as, for instance, the virtual character taking a step in the opposite direction of the collision. We could also trigger verbal reactions to express that virtual characters are embarrassed by collisions. Adding such virtual behaviors combined with haptic feedback could improve participants' immersion and feeling of presence.

Finally, one last point concerns the many devices (armbands, MSIvirtual realityone, HMD, X-Sens, etc.) required to be worn by participants for a significant amount of time. Carrying such equipment can have an effect on participants' motion as well as comfort. In our case, the experience was still relatively short and lasted only for 15 to 20 minutes. However, longer immersion durations might require to use wireless HMD solutions instead, even if this today means decreasing the field of vision.

4.5 Conclusion

In this chapter, we designed an experiment to evaluate the effects, as well as the after-effects, of haptic rendering on a motion task in a highly crowded environment. Participants performed a goal-directed navigation task through a dense virtual crowd. Wearable haptic devices provided them with vibrotactile feedback whenever a collision with their arms occurred. Results showed that providing haptic feedback impacted the way participants moved through the virtual crowd. They were more cautious about the collisions they provoked with virtual characters, but they did not change their global trajectories. We also demonstrated the presence of an after-effect of haptic feedback, since changes in their movements remained after haptic feedback was disabled. Finally, quite surprisingly, we did not notice any impact of haptic rendering on the perceived Presence and Embodiment. These results show that visual information is probably the main sense used for navigation in dense crowds. However, a combination of visual and haptic feedback improves the overall realism of the experience, as participants show a more realistic behavior: they are more cautious about not touching virtual characters. For this reason, we therefore suggest using haptic rendering to study human behavior and locomotion interactions that may lead to contacts.

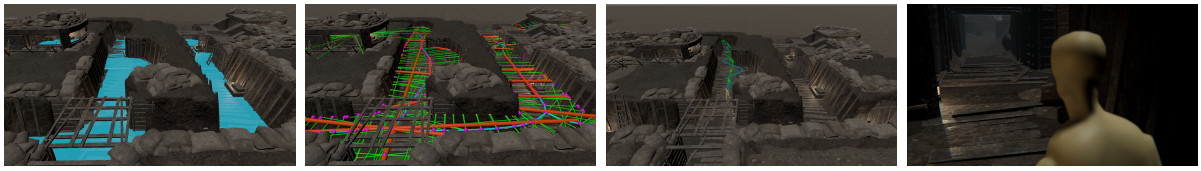
For future work, we are interested in populating our virtual environments with more interactive and reactive virtual characters. This is a crucial aspect since it seems to be a requirement to further improve the feeling of presence of participants. Also, the use of reactive characters may increase the effect of haptic rendering, since we could expect stronger participant reactions when virtual characters would also react after a collision. A more detailed analysis that evaluates motion before and after a collision is rendered and a virtual character reacts would then also be relevant to study. We are also interested in carrying out experiments enrolling more subjects and analysing a wider range of metrics in different scenarios (e.g., considering a dynamic crowd, measuring the effect on shoulder hunching, carrying out a control experiment where no haptics is applied). Finally, as we mention in the limitations, we plan to use more compelling wearable haptic devices to provide a more realistic sensation of collision while keeping the overall system compact and easy to wear, e.g., skin stretch [Chinello et al., 2017a] or tapping devices for the shoulder and upper arm.

Chapter 5

Virtual camera control

Contents

In this chapter we present our last contribution. In the previous chapters we described how to generate, study and perceive various non-verbal characteristics of the motion, in this one we focus on how to convey those to an audience. In the Introduction, we introduced why conveying specific motion characteristics to an audience requires specific rules, which typically requires to define a cinematographic language [Arijon, 1991]. Indeed, a key component of audience access to video production is how the camera is placed and moved according to scene contents, agent motions as well as narrative and stylistic constraints. These requirements are shared between all the visual experiences. Even for video games, in which cameras are directly controlled by the players, there are pressing requirements to create well-shot cinematographic sequences from single or multiple playing sessions that could then be streamed to larger audiences, typically for e-sports games. In these cases, with no specification and no previous knowledge on the events, an additional requirement is the ability to decide at run time, while the user/s is/are playing, the best



(a) Environment and navigation mesh (b) Auto-generated camera tracks (c) Actor and camera trajectory (d) View from the camera

Figure 5.1 – Our topology-aware camera control system works as follows: starting from a virtual environment with its navigation mesh in blue (a), a collection of camera tracks are generated by clustering points obtained via ray casts (green) generated from a topological skeleton representation of the navigation mesh (b). The camera is then controlled in real-time by a physical system that follows a target on the best camera track in order to film an actor navigating in the environment (c and d).

camera angles and displacements which satisfy narrative and stylistic visual constraints in complex 3D environments.

To date, most cinematographic camera systems rely either on (i) the prior manual placement of cameras by artists in 3D environments which are then triggered at runtime by events in the game (*e.g.* characters entering a building, climbing stairs, jumping between platforms), or (ii) the use of motion planning techniques through the computation of camera roadmaps in the 3D environment (*e.g.* probabilistic roadmaps [Li and Cheng, 2008; Nieuwenhuisen and Overmars, 2004]) which are exploited at runtime but generally fail in creating a cinematographic look-and feel. Automated placement of multiple cameras has also been addressed in the specific case of designing staging and shooting layouts. For instance, Louarn et al. [2018], the authors rely on a high-level specification language to place both the camera and the characters in relation with the environment. The work deals with the optimization of event visualization, but the main difference is that here the positioning of the cameras is in relation with the positioning of the agents, and both tasks are addressed at the same time. In our work we have no information of where the agents will be, so we have to rely on hypotheses as to where the characters will be and how they will move, and ensure that there are enough cameras to cover their range of positions/motions. In this chapter, we present work that focuses on camera placement in a known environment to follow real-time events, targeting use cases such as e-sports and video game streams, conveyed in a cinematic fashion.

Challenges. Creating such a system requires addressing the following challenges (i) automatically create camera angles and camera tracks of cinematographic quality (ii) con-

necting camera angles and camera tracks together in a joint representation to enable continuous or discrete transitions and (iii) at run-time, computing the camera motion and cuts given a number of targets to follow and high-level constraints (static *vs.* dynamic cameras, shot sizes, anticipation *vs.* lazy cameras, cutting pace). Addressing the problem first requires a better understanding of underlying motivations and constraints which guide the design of camera in real movies and endow them with a cinematographic look-and-feel. A first observation is that this design is predominantly a matter of directorial style. For the same motion of characters, there are significant variations in how the cameras can be placed and moved [Jiang et al., 2020]. Therefore, an artistic control over the camera parameters is required. A second observation is that camera tracks are strongly driven by the topology of the environment. For example, when considering the design of a camera sequence in a corridor, there is little number of alternatives in trajectories: motions all follow the shape of the corridor, generally in a close to linear motion where possible, tracking characters from front, side or rear view.
In addition, linear or close-to-linear camera motions are prevalent in real movies, first due to physical constraints of camera rigs (mostly linear rails or camera dolly carts), and second due to their visual simplicity (complex motions tend to distract the spectator from the content, unless it is the intention). At last, static cameras are commonplace in movies. When tracking characters, these cameras are placed at locations which maximise visibility, and generally pan to follow characters motions (unless implementing specific intentions such as cameras placed behind hedges to enforce partial visibility).

We noted the following requirements:

— populating the environment with static cameras observing large areas.
— populating the environment with linear camera motions that simulate classical dolly track motions.
— populating the environment with a network of linked camera paths which would enable following a character without cuts whatever the motion it performs.

Contributions. To address these requirements, we propose a *topology-aware* approach designed in two phases. A first offline phase that exploits navigation meshes in 3D gaming environments to build a simplified skeletal representation. Omni-directional or controlled directional ray-casts are then performed from the skeletal representation to the scene geometry, to populate the environment with virtual cameras along the scene geometry and aiming at the skeleton. Virtual cameras are then clustered using sequential RANSAC

with a linear model to extract pieces of linear camera motions. Finally, all linearized motions are linked in a graph representation. A second and online phase that computes at each frame a virtual target position on the edges of this graph, representing an optimal camera position and then a physical camera model is used to attract the virtual camera towards the optimal camera.

Our contributions are threefold:

— a novel approach to automatically compute a collection of camera angles and camera tracks which are aware of the scene topology and implement different directorial styles, using a sampling+clustering approach;

— a graph representation dedicated to camera control: the *camera navigation graph* which abstracts the regions in which the camera can move, enables efficient queries, and yields smooth camera motions;

— a real-time cinematographic system which can compute in real-time (in less than 20ms), smooth camera motions and automated transitions between viewpoints, responding to high-level directorial constraints (camera distance, camera angle, cutting speed, static or dynamic tracking)

As a result, this opens many possibilities for real-time fully automated cinematographic systems deployed in game engines with complex environments and interactive control of directorial style, such as in e-sports live casting events, where game sessions can be conveyed in more cinematographic ways and display characters' motion characteristics by borrowing and adapting techniques from real movies.

The chapter is structured as follows. We start by introducing the overview of our approach in Section 5.1. The two main parts of our approach are then detailed in Section 5.2 (offline pre-processing stage) and Section 5.3 (real-time camera placement). Section 5.4 presents some artistic results and comparisons with probabilistic roadmaps. Finally, the discussion and future works are presented in Section 5.5.

5.1 Overview

As mentioned, our approach provides a real-time generation of cinematic cameras in game-like environment, through two stages: The offline pre-processing stage (detailed in Section 5.2) takes as input a *navigation map*, *i.e.* a 3D topology which represents the surface on which characters can navigate in 3D environments. A geometric skeleton is

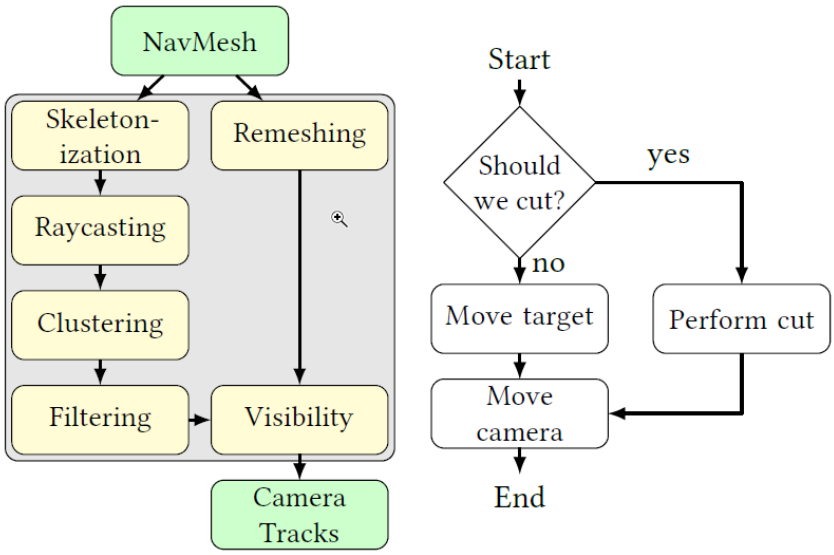


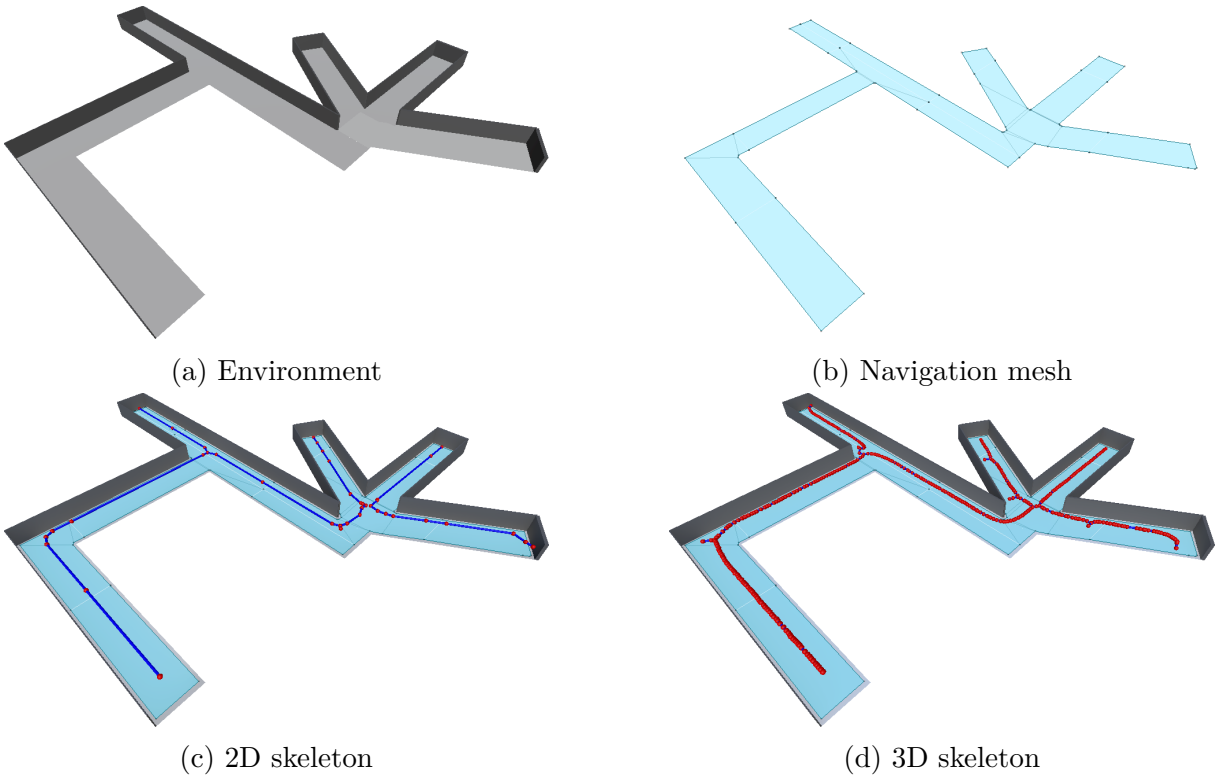
Figure 5.2 – Overview of our system.

extracted from the topology to provide an abstract and simplified representation of the navigation map. The skeleton is then used as a baseline on which (i) a raycast sampling is performed, by shooting rays locally orthogonal to the skeleton towards the 3D environment. Hits of the rays and samples from the skeleton compose a collection of camera poses. Then with a sequential RANSAC process we perform a multi-model estimation, where our model is linear pieces of camera motions. Linear motions are further cleaned, and structured into a camera navigation graph.

The second process, detailed in Section 5.3, uses the camera navigation graph to decide in real-time where to place and how to move the camera according to the position of an entity. Designers can tune some elements such as framing and cutting strategies to influence the camera placement in real-time.

5.2 Precomputation

The first stage of our system consists in an offline computation, the input of which is a navigation mesh (*navmesh*) *i.e.* a 3D triangulated polygon, subset of the environment. A *navmesh* is a common representation used in 3D applications for navigation agents, which encodes the surfaces on which the agents can move. Navmeshes are supported by all mainstream 3D game engines (such as Unity and Unreal Engine). As displayed in Figure 5.5, this process is separated in six distinct steps, which are later described.



(a) Environment (b) Navigation mesh

(c) 2D skeleton (d) 3D skeleton

Figure 5.3 – 2D and 3D skeletonization of a navigation mesh without overlaps in height results in different skeleton representations.

Skeletonization. A skeleton [Aichholzer et al., 1996; Brandt and Algazi, 1992] of the navigation mesh is extracted: it provides an abstraction of the topological characteristics of an arbitrary environment (*e.g.* corridors, intersections, forks, dead-ends).

Raycast sampling. A sampling process using raycasts from the skeleton to the 3D environment along heuristic directions that creates a cloud of possible camera positions either along the environment (if the rays hit the environment) or in mid-air (to a cut-off distance if there is no hit).

Clustering. A clustering of the possible cameras is performed using a multi-model fitting algorithm (here a sequential RANSAC [Fischler and Bolles, 1981] for its $O(n)$ performance) to extract a collection of underlying linear sections which will become camera tracks.

Filtering. A filtering stage is performed to remove specific artifacts from the clustering (*e.g.* a track that collides with the environment).

Dual visibility estimation. To reduce the cost of visibility computation at run-time, we estimate the visibility between camera nodes and triangles from the navigation mesh, in a way similar to Oskam et al. [2009] using Monte-Carlo raycast sampling. To increase precision in the estimation, a mesh refinement is performed on the navigation mesh to obtain triangles under a given area [Botsch and Kobbelt, 2004]. Visibility estimation is stored both in the camera nodes (the list of triangles visible from this camera) and in the triangles (the list of cameras which see this triangle).

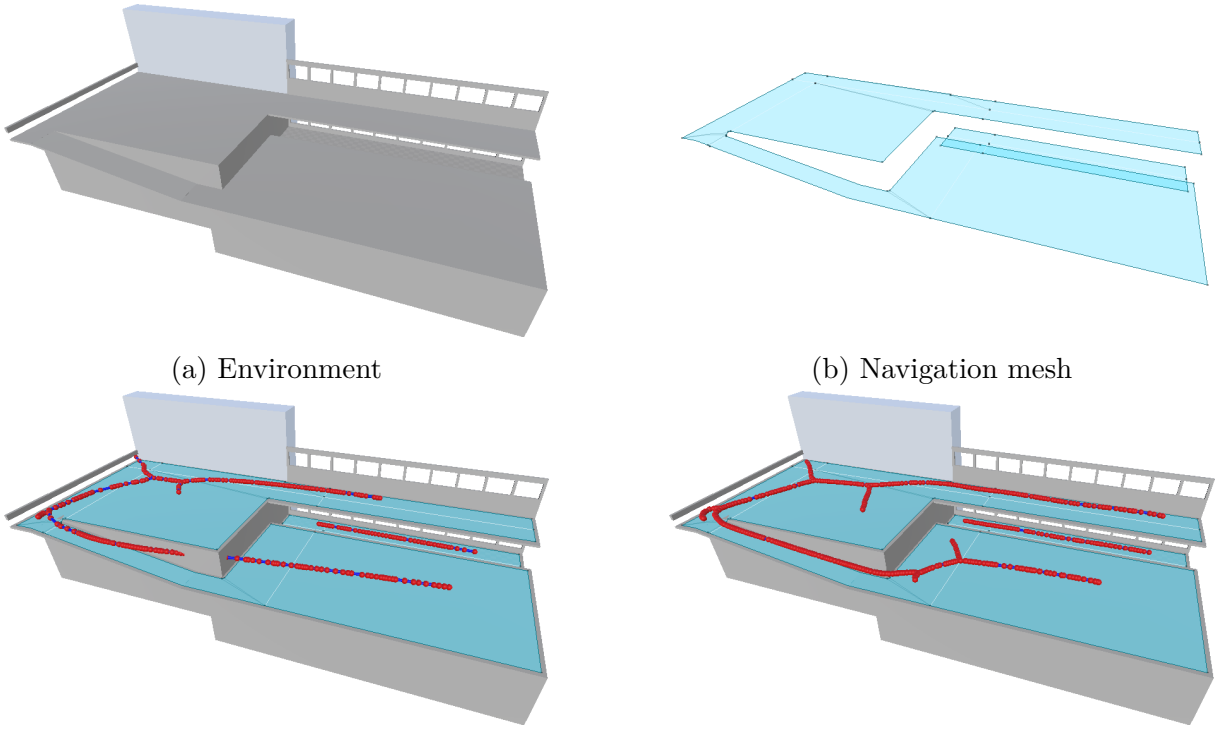
Building a camera navigation graph. The last stage finally links the isolated cameras and camera tracks into a *camera navigation graph* which can be efficiently queried to decide where to place and how to move the camera.

The output of this process is (i) a camera navigation graph representing possible camera locations (the nodes) and possible camera tracks (the edges),and (ii) the visibility information relative to a remeshed navigation surface.

5.2.1 Skeletonization

The purpose of this first stage is to obtain a simplified and abstract representation of where entities (*e.g.* characters) can navigate in a given 3D environment. The first step of the process is to extract a topological structure of the environment. We propose to rely on the navigation mesh which is an approximation of the environment that can be automatically computed [Lamarche, 2009; Oliva and Pelechano, 2011; Xiang Xu, 2011] and obviously offers a complete representation of where entities can be located. This information remains however complex to process and analyze if corridors, intersections or forks need to be detected. To both abstract and simplify this representation, we propose a topological skeleton extraction from the navigation mesh using straight skeletons [Aichholzer et al., 1996] and mean curvature skeletons [Tagliasacchi et al., 2012].

Straight skeletons were introduced by Aichholzer et al [1996]. as a replacement of widely used medial axis techniques, for its lower computational cost and simple straight-line structure. A straight skeleton is solely made of line segments which are pieces of angular bisectors of polygon edges, and computed using a shrinking process on possibly non convex polygons. Straight skeletons are limited to 2D polygons only. Therefore, for all



(a) Environment

(b) Navigation mesh

(c) 3D skeleton with 3 iterations

(d) 3D skeleton with 5 iterations

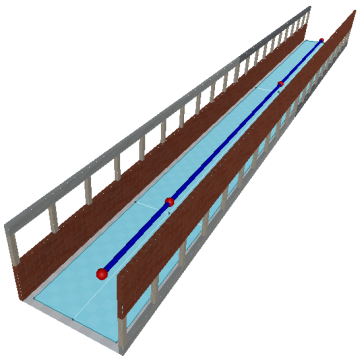
Figure 5.4 – 3D skeletonization of a navigation mesh with overlaps in height. Notice the influence of the number of edge-split iterations on the resulting skeleton: with 3 iterations (c) it intersects the environment and with 5 iterations there are no intersections (d).

navigation meshes where projection on a 2D plane does not yield overlapping surfaces, we simply (i) perform the straight skeleton extraction on the 2D projected navigation mesh and (ii) reproject the skeleton vertices to the original navigation mesh.

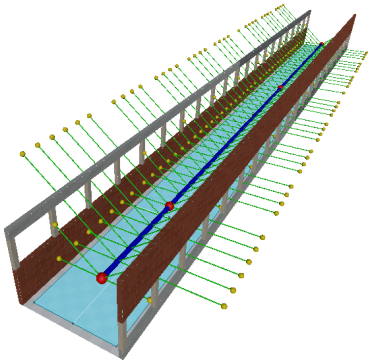
For navigation meshes where 2D projections overlap, we propose to rely on mean curvature skeletons [Tagliasacchi et al., 2012]. The mean curvature technique collapses a given 3D mesh into a skeleton structure using mean curvature flow and Voronoi medial skeleton to obtain a medially centered curve skeleton. A well-centered curve skeleton is computed by minimizing the energy function E:

$$E = E_{\text{smooth}} + E_{\text{velocity}} + E_{\text{medial}}$$

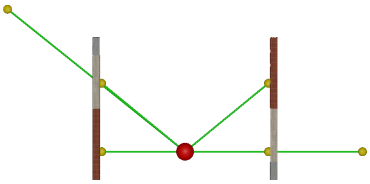
where E_{medial} energy pulls the evolving surface towards the medial axis, at an energy velocity E_{velocity} depending on the curvature, with a smoothness controlled by energy E_{smooth}. To apply this technique we (i) first extrude the navigation mesh by a height



(a) Test environment, navmesh (in cyan) and skeleton (in blue).



(b) Casted rays (green) and resulting points.



(c) Profile view. Ray directions follow different heuristics for well placed cameras (here at level or 45° angle).

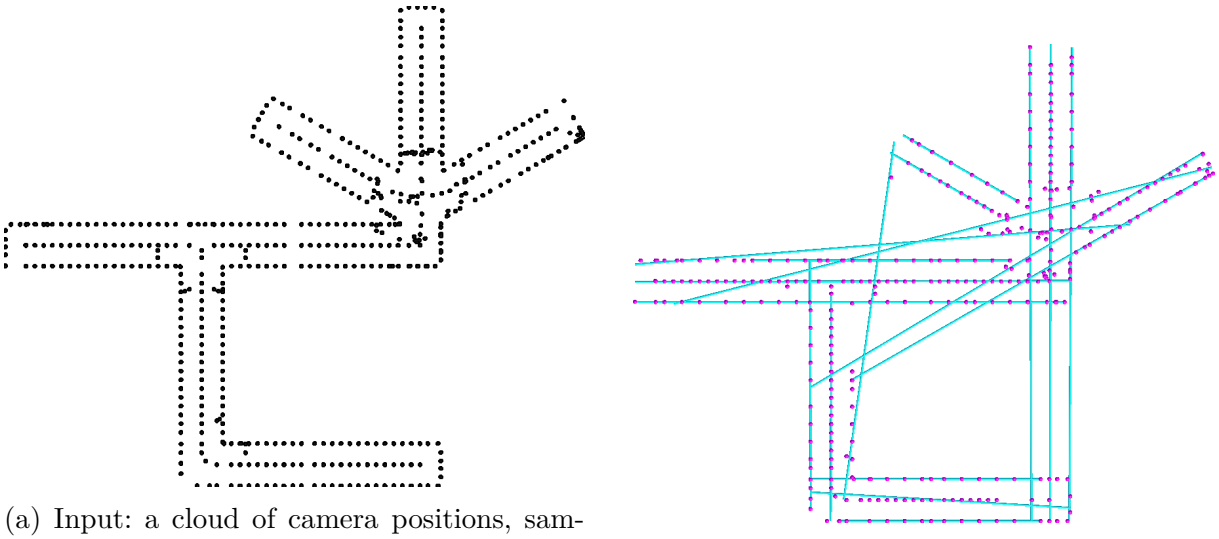
Figure 5.5 – Raycast sampling on a simple environment with walls and windows. The skeleton (red and blue), the rays (green), and the resulting points (yellow). Note how some rays intersect the environment while others go through the windows/open areas.

representing the size of an entity (typically the character) navigating on this mesh, (ii) then perform edge-split iterations to refine the mesh (as described in [2012] to improve quality) and (iii) apply the mean curvature technique.

In terms of computational cost, the straight skeleton technique is more efficient (*e.g.* 0.3s vs. 9.7s for examples presented in Figure 5.3c and Figure 5.3d). Also, while the quality of the 3D skeleton gets better with a more complex input mesh, the computational time gets higher (*e.g.* 0.4s for 3 edge-split iterations vs. 7.7s for 5 iterations in examples presented in Figure 5.4c and Figure 5.4d).

5.2.2 Raycast sampling

The skeletal representation provided in the previous stage abstracts the motion of the characters on the navigation mesh to a sequence of segments. We exploit these segments to automatically generate a large collection of cameras by casting rays orthogonal to the segments, hence towards the scene geometry since the segments represent local medial axes/mean curves. Intuitively, we are generating camera samples which follow the shape of the skeleton from far enough to provide a larger view on the overall motion of the characters. In addition, the casted rays adapt to all the geometries of the environment, including the ones not considered by the *navmesh*, creating cameras at different depths from the skeleton and through open windows/doors.



(a) Input: a cloud of camera positions, sampled from the skeleton points and raycast samples

(b) Multi-model RANSAC clustering to extract 3D line structures from camera samples.

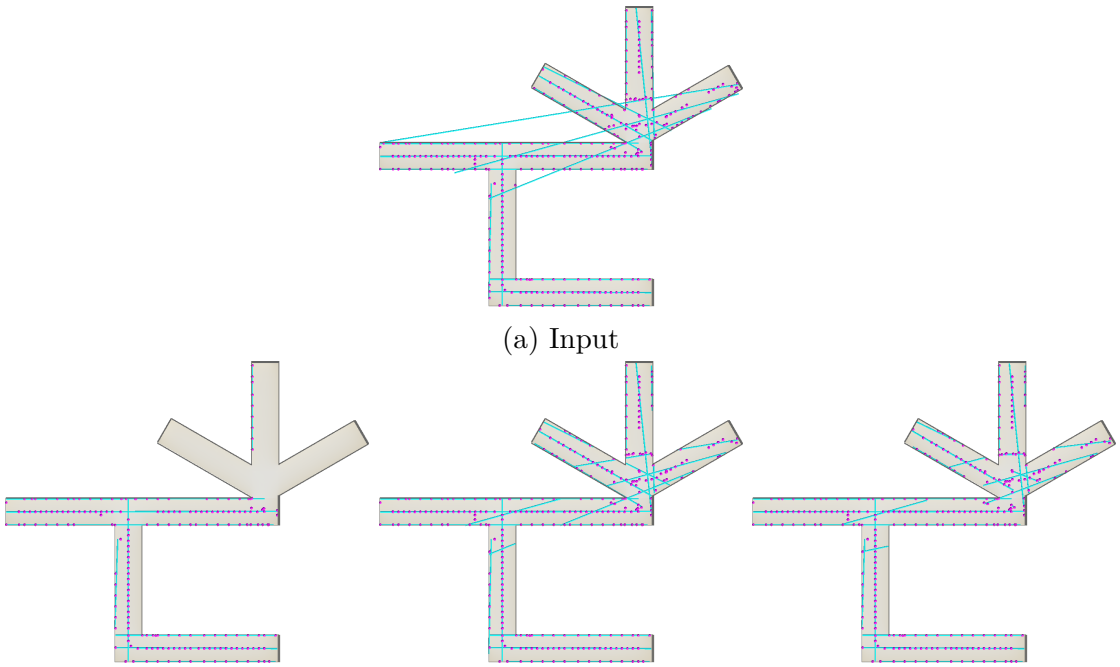
Figure 5.6 – Clustering cameras to create linear camera tracks is performed by using a sequential RANSAC process.

To compute these camera positions, we propose to cast rays from the skeleton in a number of heuristic directions that correspond to cinematographic camera angles, *e.g.*, cameras at the same height as the characters, as well as high angle, low angle or birds' eye angles (a camera above the character). If the ray intersects the environment, we will place the camera on the ray at an given ϵ offset from the environment. If the ray does not intersect the environment, a specific threshold distance $d_{\max}$ is used to bound the position of the camera on the ray.

This heuristic sampling step is meant to be flexible and personalized by the user based on the preferred styles, by choosing the directions and the distances of the rays. The result of this step is a heterogeneous point cloud of camera locations (displayed in yellow in Figure 5.5b) with points resulting from a direct projection of the skeleton towards the sides, either creating an offset of the path, or adapting the offset to the scene topology (pillars, windows, etc). Each of these cameras also encompasses the direction of its associated ray and the origin of the ray on the skeleton.

5.2.3 Clustering

The raycast sampling step computes a cloud of camera locations from which the navigation mesh skeleton is visible. In order to compute a set of camera tracks, we propose



(a) Input

(b) Filter 1

(c) Filters 2 and 3

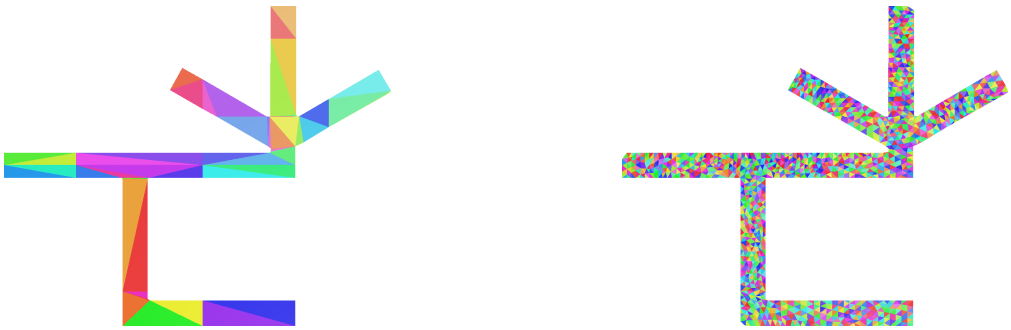
(d) Filters 2, 3 and 4

Figure 5.7 – Computed camera tracks (in cyan) are filtered to remove artifacts which occur when clustering lines from different parts of the geometry.

to cluster the camera locations using a multimodel fitting algorithm using a line model. In this way we aim to identify underlying linearities both from the skeleton and the geometry of the surrounding environment. This is a problem for which many solutions have been proposed (see [Li et al., 2017] for a detailed comparison). Here we rely on a sequential RANSAC method [Fischler and Bolles, 1981] which performed better than other approaches (multi-RANSAC, residual histogram analysis or J-linkage [Fouhey et al., 2010]) on our datasets, and is of $O(n)$ complexity. Given a model μ, RANSAC extracts a consensus set CS from a collection of points $\mathcal{P}$ such that:

$$CS(\mu, \mathcal{P}, \epsilon) = \{p \in \mathcal{P} | R(\mu, p) < \epsilon\}$$

where R is the error function. We used the standard point-to-line distance metric R as an error. All inliers of the first consensus set, *i.e.* $CS(\mu, \mathcal{P}, \epsilon)$, are removed from $\mathcal{P}$ and RANSAC is re-applied on the result until a given number of iterations is reached. As displayed in Figure 5.6 the corresponding camera tracks (in cyan) display a number of artefacts which need to be corrected through filtering.



(a) Original triangulation (b) Remeshed triangulation

Figure 5.8 – A mesh refinement stage performed on the navigation mesh as support for visibility evaluation.

5.2.4 Filtering

The obtained camera tracks are defined by their supporting points (inliers from the sequential RANSAC) and since the clustering step only takes as input a point cloud and not the geometry of the environment, the camera tracks might display a number of issues (*e.g.* collision with the environment) depending on the environment and the parameters. We propose four filters that the user can use in any combination.

Filter #1 removes the parts of the tracks which collide with the environment in order to avoid the camera moving through a wall (see Figure 5.7a). We use a iterative splitting approach to decide which segments to cut. Filter #2 removes the points which have no visibility to their associated line (*i.e.* the ray between the point and its projection on the line intersects the environment). Filter #3 removes lines without supporting points, as previous filters can leave "empty" lines after a split. Filter #4 recomputes the equation of each line using a single RANSAC step to better fit the data after a split has occurred.

5.2.5 Mesh refinement and visibility estimation

While we ensured that each camera location computed during the raycast sampling step had visibility towards a single point on the skeleton, this remains insufficient. To avoid performing visibility computation at run-time, we propose to pre-compute the camera-to-mesh visibility with the static parts of the 3D environment. We draw inspiration from Oskam et al. [2009] whom perform inter-visibility estimation for each couple of samples in the environment. To reduce the cost of the process, we only perform inter-visibility estimation from each node to each triangle of the navigation mesh, inside a limited range, using

ray-casting. Prior to visibility estimation, we perform an anisotropic remeshing [Botsch and Kobbelt, 2004] to refine the size of all the triangles of the navigation mesh (see Figure 5.8a). This improves the precision of the visibility estimation. We store the visibility estimation both in the refined triangles of the navigation mesh (each triangle knows which cameras see it) and in the camera (each camera knows the triangles it can see). The cost in terms of memory usage grows linearly with respect to the number of cameras, and for each triangle only the degree of visibility and triangle ID is stored.

5.2.6 Camera navigation graph

The last step aggregates the results of the previous steps in a data structure that can be efficiently queried at run-time to compute a camera position or motion. We propose to use a non-directed graph, where each node represents a possible camera location in the environment and each edge represents possible transitions between these locations. Each node therefore needs to encode all the data necessary to efficiently place the camera: 3D position, transitions to other nodes, and the portions of environment visible from this node. The graph is computed as follows. (i) First, each camera track from the filtering step is inserted in an arbitrary order in the graph. Two nodes are created for the endpoints of a camera track, and an edge is created to link the two end points. Then, new nodes are created at the intersection between newly created edges and existing ones. This enables tracks interconnection. (ii) Second, edges are split by inserting new nodes so that each edge is shorter than a user-specified length. (iii) Third, strongly connected components in the graph are linked together by linking nodes that are closer than a user-specified threshold while ensuring visibility. This enables camera tracks to easily connect to their neighbor tracks, hence creating a camera navigation graph. (iv) Lastly, points corresponding to each node are inserted into a KD-tree in order to accelerate run-time queries.

5.3 Real-time camera placement

In this stage we implement a simple camera placement system to illustrate the features of the camera navigation graph. We rely on the Unity's Cinemachine framing system (which smooths jitter movements with a dampened mass-spring system) and on its virtual cameras system to reduce the impact of managing a potentially unlimited number of cameras that our system can generate (each framing a subject, that may be shared with

other cameras). The inputs of our system are, for each camera, (i) a subject to frame, along with its height; (ii) a framing strategy, dictating how the subject should look on the screen; (iii) a movement strategy, dictating how the camera should move in the environment; and (iv) a cutting strategy, specifying the conditions under which a cut should be performed. To define the camera position, the system computes, at each frame, a *target* on the tracks that represents the current best possible position, given the specified strategies. Then, the actual camera position is computed by using a physical system in which a force is attracting the camera towards the target. This system, similarly to a low-pass filter, helps reducing jerkiness of the movements.

Each iteration of our algorithm, *i.e.* a frame, comprises the following steps. First, a *cutting strategy* algorithm evaluates whether a cut needs to be performed (see Section 5.3.1). In such case, a new target position on the tracks is computed following classical continuity rules (see Section 5.3.3). If no cutting is required, the target position is updated on the track using a *target moving strategy* (see Section 5.3.2). Once the target position is updated, the camera is moved using the force-based system. Lastly, the camera position is updated.

5.3.1 Cutting strategy

In order to avoid an unnecessary and expensive search for a new target position, a number of checks are performed. These checks are all the ones that do not need an updated target position, and each of them can be controlled by the user as part of the cutting strategy. There are three conditions which may trigger a cut:

— shot duration with a log normal distribution model [Galvane et al., 2015b];
— visibility check, through raycasts, to ensure the subject is not occluded for more than a given duration (200ms) by a static or a dynamic obstacle.
— framing quality, evaluating if the user-specified shot size (character on the screen) is not violated for more than a specified duration(200ms).

5.3.2 Target moving strategy

The target always moves on the camera tracks (the edges of the camera navigation graph). To find the optimal position for the target, we first need to select the appropriate edge, then find the right position on that edge. As the position of the target cannot be predicted too far ahead in time, the selection of the best target on the camera navigation

graph is not straightforward.

We propose the following algorithm. First, a set of edges is gathered by iterating on the closest to the actor (*i.e.* edges connected to nodes that are closer than a user-specified distance). Unwanted edges (such as those not strongly connected to the previous edge) are filtered out. Then we identify potential point of interests (POI) on these edges: projection of the user on the edge line, points at the right framing distance from the actor.

Next all the POI are scored. This score is the weighted average of 4 sub-scores, with user-specified weights that constitute the framing strategy. The first sub-score is the *shot size*: using the vertical field-of-view angle of the camera and an expected on-screen height of the actor, an optimal distance camera-actor is computed. The second sub-score is the *direct visibility*, making sure that the actor is still visible by casting a ray between the POI and the actor, and monitoring if this ray intersects the environment. The third sub-score is the *indirect visibility* that tries to assess how much of the actor's surroundings are visible from the POI. This score is computed by first identifying triangles from the remeshed navmesh (see Section 5.2.5) around the actor, and computing the percentage of those that are visible from the POI using rays. The last sub-score is the *distance*, making sure that the new target position is the closest to the camera. This score is computed using the graph distance, the shortest path on the track between the two points (computed using an A* algorithm).

The new target position is then selected from the POI with the better score using a gradient descent by moving on the graph edges around the POI by fixed intervals to find the local minimal score. If, for any given reason, no point of interest can be found, a cut is needed.

5.3.3 Continuity rules

A cut is computed in a similar way as the target presented in Section 5.3.2 except that no edge filtering is performed. The score for each POI is computed using three of the four previous sub-scores (shot size, direct visibility and indirect visibility) and two additional cut-specific sub-scores. The first score, the 30◦ rule, is derived from a classical cinematographic rule [Arijon, 1991] stating that the angle between the pre-cut camera, the actor, and the post-cut camera must be over 30 degrees to avoid jump-cuts that distract the spectator. This score is computed by using the cosine of the angle between the projection of the actor's velocity vector on the pre-cut camera and its projection of the post-cut camera. The second score, the *optical flow*, is derived from the "line of action"

	Pre-computation	
	Mean	Deviation
skeleton	6.45s	2.65s
remeshing	0.04s	0.01s
raycasting	0.02s	0.01s
clustering	4.90s	10.85s
filtering	0.02s	0.01s
tracks	0.84s	0.90s

Table 5.1 – Average time for 222 pre-computation on the environment

rule, saying that during a cut, the camera should not cross the line of action, so that the actor's movement, seen by the camera between the cuts, have similar directions. It is computed using the cosine of the angle between the projection of the actor's velocity on the pre-cut camera and its projection on the post-cut one.

5.3.4 Moving the camera

Once a new target position is computed, we can move the camera towards it. We have two possibilities. (i) The target is the result of a cut, then the camera can "jump" directly to the target position, and be oriented in the direction of the actor. (ii) The target is not the result of a cut, then we use a force-driven system, with two forces: one attracting the camera to the target, and a second one repulsing it from the actor. Therefore, the acceleration of the camera is the weighted combination of these two forces, with a mass set to 1 kg to avoid "overshooting" the target and so an unpleasant back-and-forth motion of the camera. The user can also define a maximum velocity.

5.4 Results

We show the relevance of our approach by studying a realistic scenario. All computations were done on an Intel Core i7-9850H laptop at 2.60GHz with 32GB of memory.

The environment elected for the scenario is a First World War-inspired trench scene available on the Unity Asset Store [1]. This environment is composed of two distinct sets of trenches (see Figure 5.9) and the navigation mesh is composed of 473 triangles and 1017 vertices. Times for each pre-computation step is shown in Table 5.1.

1. https://assetstore.unity.com/packages/3d/environments/historic/world-war-trenches-152381

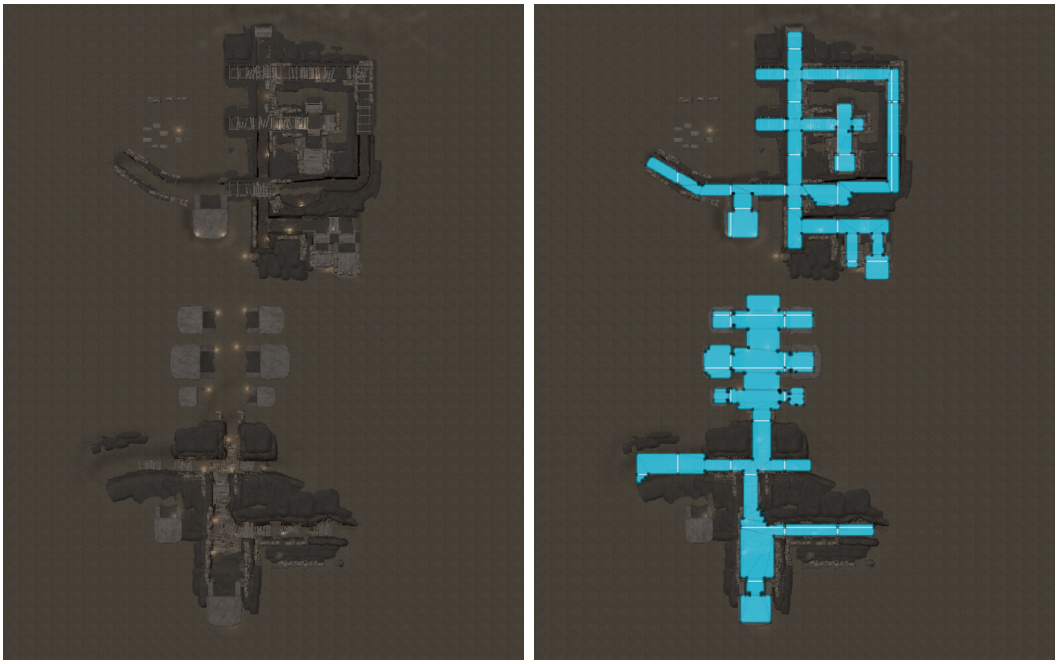


Figure 5.9 – Trench environment (left) and corresponding navigation mesh (right) used as a benchmark scenario.

5.4.1 Artistic control

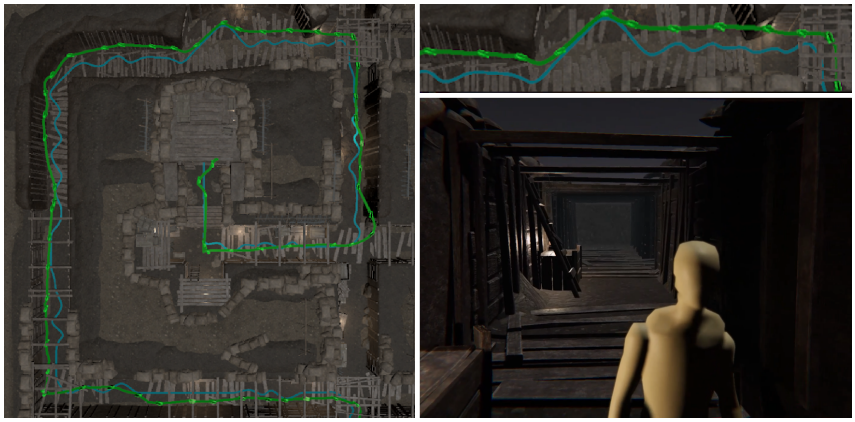
We provide a set of artistic tools to control the camera. Here we present the different styles that can be achieved using the same pre-computation step. Shown in Figure 5.10 are three different camera controls in framing size (Medium shot, Long shot and Extreme long shot) and in allowed camera movement: either dynamic with no cuts, dynamic with cuts and static cameras. The trajectory of the actor (in cyan) is the same in all videos.

Shown in Figure 5.11 is the influence of the parameters during the pre-computation on the camera angles and framings that can be obtained at run-time. Please refer to the supplementary video for the full-length videos of all these scenarios [2].

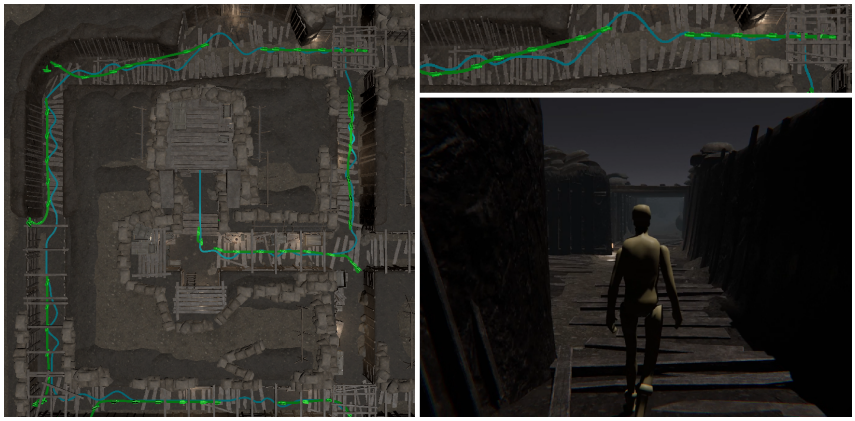
5.4.2 Comparison against Probabilistic Roadmaps

We compare our graph generation approach to a probabilistic roadmap (PRM), which is a technique that generates a motion graph by randomly sampling a number of points in the environment, and linking each pair of points if the arc does not intersect the

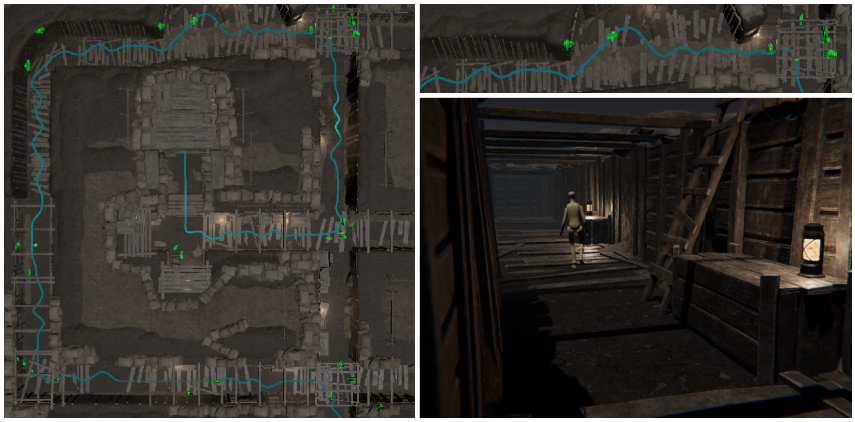
2. https://youtu.be/9UQS9_84F70



(a) Medium shot, cuts prohibited



(b) Long shot, cuts allowed



(c) Extreme long shot, static cameras

Figure 5.10 – Outputs with different parameters: trajectories (on the left) of the actor (in cyan) and the camera (in green). Examples of camera frames are also provided on the right.

(a) Raycasting only upwards

(b) Raycasting on the sides and upwards

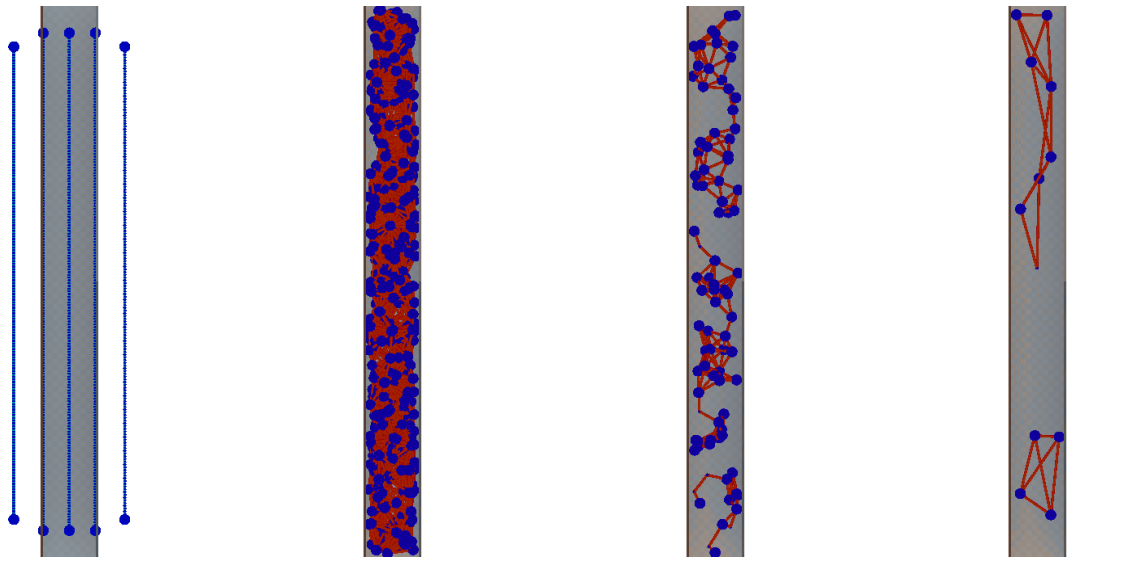
Figure 5.11 – Influence of the track generation on the obtained images. Actor trajectory and position are the same in both pictures.

		Target update		Camera update	
		Mean	Std. Dev.	Mean	Std. Dev.
Cutting	Ours	0.01s	0.04s	0.000 09s	0.000 48s
	PRM	0.05s	0.02s	0.004 78s	0.004 44s
Not cutting	Ours	0.02s	0.03s	0.000 01s	0.000 01s
	PRM	0.19s	0.10s	0.000 00s	0.000 02s

Table 5.2 – Cost of positioning the camera (per frame).

environment. We compare the two techniques on the same environment, using the same camera position algorithm described above, and having the actor take the same path.

To compare these two techniques, a simple metric is to compare the time needed to compute a camera position per frame. As shown in Table 5.2, our method is on average 5 times quicker when looking for a new target position. This time difference is mainly due to the difference in arity between the generated graphs (*e.g.* 381 nodes and 1272 edges with our technique, 1851 nodes and 22786 edges with PRM on the same environment). This difference can be explained by the fact that while our technique tries to only generate tracks that are cinematographically interesting, PRM creates points at random, therefore needing a higher number of points for a correct result.



(a) Ours (b) Dense PRM (c) Sparse PRM (d) Far PRM

Figure 5.12 – Visual comparison of the camera tracks. The environment is a simple corridor with windows (as in Figure 5.5). In (b) point density: 1.5, link distance: 5. In (c) point density: 0.1, link distance 5. In (d) point density: 0.01, link distance 20.

This effect is highlighted in Figure 5.12 with a toy environment composed of a single corridor with windows. Even without taking into account the fact that our method generates tracks outside the corridor that view the inside through the windows thanks to the raycasting step, it is clear that the tracks generated via a PRM are not as straight or useful for placing a camera. If we test PRM with a number of nodes comparable with the one generated by our techniques the PRM generates sub-graphs that do not span the entire length of the corridor, and this generates blind areas with no camera coverage, in larger and more complex environments. On the contrary, if the density of points is too high, then the entire environment is covered and a camera moving on the graph is akin to a free camera, which defeats the purpose of creating camera tracks.

5.5 Discussion and future works

This contribution addresses the problem of automatically populating 3D environments with static cameras and linear camera rails. By analyzing the navigation mesh of 3D environments, we designed different camera placement strategies which are based on an abstract representation of the environment and exploit the topology of the environment.

We individually demonstrate the relevance of these strategies on a number of illustrative examples, and display results with all strategies on a large and complex 3D environment.

The camera system we built is based on the prediction of potential displacements of agents in a complex virtual environment, conveying a cinematographic experience to the audience.

In the future, we could extend our approach with new camera generation strategies and a high level control, as well as by making available a wide range of cinematographic styles, yielding the ability for the director (should it be virtual or real) to vary in style depending on the context, the events and atmosphere to convey. The focus of our work was on anticipate agents displacement in a known environment, we believe that, such systems, could profit from techniques to interpret and react to additional non-verbal characteristics of of the characters actions, like by triggering specific style relate to detected communication clues(*e.g.* if the character is friendly or aggressive).

Conclusion

Providing virtual humans with non-verbal communication capabilities and exploiting them in interactive applications is a challenging problem, that involves multiple domains such as animation, psychology and sociology. In this manuscript, we presented a transversal exploration on non-verbal interactions with virtual humans, through the contributions organized around four main research axes. **Research axis 1:** *to propose efficient animation methodologies in order to convey non-verbal characteristics of virtual human motion to the users.* As part of this axis, we developed a technique for upper-body motion editing accordingly to the observer's perception of this motion. **Research axis 2:** *to understand which factors affect the communication and how they are perceived by the users.* In this context, we conducted two user studies, to evaluate the effect of upper-body editing and gaze behaviors on the user perception. **Research axis 3:** *to study technologies that simulate the flow of sensory information from the virtual world to the user.* In the development of this axis, we designed a study to test the effects of haptic rendering for contact simulation on user's navigation through a virtual crowd. **Research axis 4:** *to propose technical solution in order to convey real-time events, involving virtual humans, as a cinematographic experience to an audience.* In this regard, we developed a tool to generate virtual cinematographic cameras positions and tracks and a system to follow the actions and interactions of virtual characters in a virtual environment.

In Chapter 2, we introduced the main contribution of this work. We developed a new paradigm, a technique for guiding agents' animation based on the observer's perception of apparent movements. We demonstrate the technique's applications across several use cases, showcasing its versatility. Furthermore, as part of the second research axis, we validate how this approach enhances users' understanding of virtual agents' intentions in virtual reality. Although we validate the prominent role of upper body motion has during non-verbal exchange, still eyes and gazes seems to have a higher perceived impact. We commented this in Section 2.3.4. Consequentially, the second study (Chapter 3) focuses on virtual human's gaze. We confirmed the presence of the stare-in-the-crowd effect

(a well-known psychological phenomenon that investigates the perceptual differences between directed and averted gazes) in virtual reality. With this result, we contribute in proving the relevance of gaze behaviors in the interactions with virtual human, and how perceptual effects we experience in the real world could similarly translate in virtual ones. The third Chapter 4 proposes a study on dynamic interaction with a virtual crowd, where virtual reality users navigate the crowd in a simulated train station, to test the effects of contact simulation through vibrotactile haptic feedback. Contact, including the violation of one's personal space, is a relevant non-verbal aspect of human interaction. However, simulating contact virtually without using obtrusive haptic devices is challenging. In our work, we tested a wearable device providing a vibrating feedback to the user's arms. The presented results demonstrate that such simulation of contacts improves the user's awareness of their personal space and reduces the frequency of collisions with virtual agents. Regarding the last research axis, our final contribution (Chapter 5) proposes a technical approach to generate cinematographic camera tracks and positions, in a virtual environment, for real-time visualization of virtual character actions. We identify potential paths for the interacting characters and evaluate visibility in these areas. At run-time, we propose an automatic system that follows the character movement accordingly to user-defined parameters.

All these results emphasize the importance of conducting a comprehensive exploration, that transverse several non-verbal characteristics related to virtual humans' perception, and demonstrates how to guide this perception in situations of active and passive interactions. Indeed, communication with virtual humans plays a central role in multiple applications, such as entrainment (movie and video games), immersive experience or as virtual assistants. We believe that this role will become even more relevant in the near future with the development of the "meta world" and increased social interactions in the virtual world.

Future works.

While we commented specific future applications for each contribution in the conclusion of each chapter, this chapter provides a general overview of how the combined knowledge of these works can be used to generate more realistic, interactive and entertaining applications involving virtual humans.

The main achievement of our investigation is placing the observer in a central role for the synthesis of virtual character motions. In the near future, we aim to further develop this concept and provide virtual humans with additional non-verbal communication functionalities, such as believable gazing patterns and more reactive motions for the upper body and displacement. We believe that such approach would help generate virtual agents with believable reactive behaviors and demonstrate awareness toward the user. Those characteristics are particularly important for immersive application, like virtual reality, where the user becomes part of the virtual world. Regarding immersive experience, we also investigate how haptic technologies enhance the way we perceive ourselves in relation with virtual humans. We only scratched the surface of what can be simulated, and how contact can play a prominent role in communication in virtual reality, including simulating natural interactions like touching one's shoulder to get their attention or handshaking. Different technologies, such as haptic vests [Raisamo et al., 2022], can be employed to achieve these goals. At last, non-verbal communication characteristics are not only relevant for the active interactants, but also for external observers. Actors train in expressive capabilities, both verbal and non-verbal, to convey emotions to an audience, working in synergy with directors to frame them within a visual narrative. For this reason, we believe that this synergy should be developed in the virtual world as well, involving animated characters that are aware of how they are observed and smart framing systems that position and move cameras accordingly. This evolution will be a crucial for the fruition of virtual events, with the growing demand in field such as e-sports.

As we have seen, this thesis paves the way for a wide range of future challenges. We elaborate this work with the belief that virtual humans will be a relevant part of our everyday lives in the future. We will interact with completely autonomous humans, but we will also embody ourselves in digital avatars. In the latter case, technology will be required to interpret our intention, emotions and behaviors and then portray them into our digital avatars. Therefore, we believe that the coordination of the design of technical solutions and the study of perceptual effects through different means will be fundamental.

Bibliography

Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., & Chen, B., (2020), Unpaired motion style transfer from video to animation, *ACM Transactions on Graphics (TOG)*, *39 4*, 64–1.

Achenbach, J., Waltemate, T., Latoschik, M. E., & Botsch, M., (2017), Fast generation of realistic virtual humans, *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, 1–10.

Adistambha, K., Ritz, C. H., & Burnett, I. S., (2008), Motion classification using dynamic time warping, *2008 IEEE 10th Workshop on Multimedia Signal Processing*, 622–627, https://doi.org/10.1109/MMSP.2008.4665151

Aggravi, M., Pausé, F., Giordano, P. R., & Pacchierotti, C., (2018), Design and evaluation of a wearable haptic device for skin stretch, pressure, and vibrotactile stimuli, *IEEE Robotics and Automation Letters*, *3 3*, 2166–2173.

Ahmed, F., Bari, A. H., & Gavrilova, M. L., (2019), Emotion recognition from body movement, *IEEE Access*, *8*, 11761–11781.

Aichholzer, O., Aurenhammer, F., Alberts, D., & Gärtner, B., (1996), A novel type of skeleton for polygons. *In J. ucs the journal of universal computer science* (pp. 752–761), Springer.

Al-Asqhar, R. A., Komura, T., & Choi, M. G., (2013), Relationship descriptors for interactive motion adaptation, *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 45–53, https://doi.org/10.1145/2485895.2485905

Allmendinger, K., (2010), Social presence in synchronous virtual learning situations: the role of nonverbal signals displayed by avatars, *Educational Psychology Review*, *22 1*, 41–56.

Arijon, D., (1991), *Grammar of the film language*, Silman-James Press.

Arikan, O., (2002), Synthesizing constrained motions from examples, *ACM Trans. Graph.*, *21 3*, 483–490.

Arikan, O., & Forsyth, D. A., (2002), Interactive motion generation from examples, *ACM Transactions on Graphics (TOG)*, *21 3*, 483–490.

Aristidou, A., Chrysanthou, Y., & Lasenby, J., (2016), Extending fabrik with model constraints, *Computer Animation and Virtual Worlds*, *27 1*, 35–57.

Aristidou, A., & Lasenby, J., (2011), Fabrik: a fast, iterative solver for the inverse kinematics problem, *Graphical Models*, *73 5*, 243–260.

Aristidou, A., Lasenby, J., Chrysanthou, Y., & Shamir, A., (2018), Inverse kinematics techniques in computer graphics: a survey, *Computer Graphics Forum*, *37 6*, 35–58.

Arnaldi, B., Fuchs, P., & Tisseau, J., (2003), Chapitre 1 du volume 1 du traité de la réalité virtuelle, *Les Presses de l'Ecole des Mines de Paris*, *1*, 131.

Ashraf, G., & Wong, K., (2000), Dynamic time warp based framespace interpolation for motion editing., *Proceedings of the Graphics Interface 2000 Conference*, 45–52.

Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., & Turk, M., (2005), Transformed social interaction, augmented gaze, and social influence in immersive virtual environments, *Human communication research*, *31 4*, 511–537.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M., (2001), Equilibrium theory revisited: mutual gaze and personal space in virtual environments, *Presence: Teleoperators & Virtual Environments*, *10 6*, 583–598.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M., (2003), Interpersonal distance in immersive virtual environments, *Personality and social psychology bulletin*, *29 7*, 819–833.

Baker, S. R., & Edelmann, R. J., (2002), Is social phobia related to lack of social skills? duration of skill-related behaviours and ratings of behavioural adequacy, *British Journal of Clinical Psychology*, *41 3*, 243–257.

Beebe, S. A., (1974), Eye contact: a nonverbal determinant of speaker credibility, *Communication Education*, *23 1*, 21–25.

Bente, G., Krämer, N. C., Petersen, A., & de Ruiter, J. P., (2001), Computer animated movement and person perception: methodological advances in nonverbal behavior research, *Journal of Nonverbal Behavior*, *25 3*, 151–166.

Bente, G., Rüggenberg, S., Krämer, N. C., & Eschenburg, F., (2008), Avatar-mediated networking: increasing social presence and interpersonal trust in net-based collaborations, *Human communication research*, *34 2*, 287–318.

Berton, F., Olivier, A.-H., Bruneau, J., Hoyet, L., & Pettré, J., (2019), Studying gaze behaviour during collision avoidance with a virtual walker: influence of the virtual reality setup, *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 717–725.

Bimbo, J., Pacchierotti, C., Aggravi, M., Tsagarakis, N., & Prattichizzo, D., (2017), Teleoperation in cluttered environments using wearable haptic feedback, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3401–3408.

Birdwhistell, R. L., (2010), *Kinesics and context: essays on body motion communication*, University of Pennsylvania press.

Blakemore, S.-J., & Decety, J., (2001), From the perception of action to the understanding of intention, *Nature reviews neuroscience*, *2 8*, 561–567.

Bodenheimer, B., Rose, C., Rosenthal, S., & Pella, J., (1997), The process of motion capture: dealing with the data. *In Computer animation and simulation'97* (pp. 3–18), Springer Vienna.

Bonneaud, S., Rio, K., Chevaillier, P., & Warren, W. H., (2012), Accounting for patterns of collective behavior in crowd locomotor dynamics for realistic simulations. *In Transactions on edutainment vii* (pp. 1–11), Springer.

Bönsch, A., Radke, S., Overath, H., Asché, L. M., Wendt, J., Vierjahn, T., Habel, U., & Kuhlen, T. W., (2018), Social vr: how personal space is affected by virtual agents' emotions, *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 199–206.

Botsch, M., & Kobbelt, L., (2004), A remeshing approach to multiresolution modeling, *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 185–192.

Boucaud, F., Pelachaud, C., & Thouvenin, I., (2021), Decision model for a virtual agent that can touch and be touched, *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*.

Bouchard, D., & Badler, N., (2007), Semantic segmentation of motion capture using laban movement analysis, *International Workshop on Intelligent Virtual Agents*, 37–44.

Bourgaize, S. M., McFadyen, B. J., & Cinelli, M. E., (2020), Collision avoidance behaviours when circumventing people of different sizes in various positions and locations, *Journal of Motor Behavior*, 1–10.

Bourgeois, A., Badier, E., Baron, N., Carruzzo, F., & Vuilleumier, P., (2018), Influence of reward learning on visual attention and eye movements in a naturalistic environment: a virtual reality study, *Plos one*, *13 12*, e0207990.

Bowen, C. J., (2013), *Grammar of the shot*, Routledge.

Brand, M., & Hertzmann, A., (2000), Style machines, *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 183–192, https://doi.org/10.1145/344779.344865

Brandt, J. W., & Algazi, V., (1992), Continuous skeleton computation by voronoi diagram, *CVGIP: Image Understanding*, *55 3*, 329–338.

Bruce, N. D., Wloka, C., Frosst, N., Rahman, S., & Tsotsos, J. K., (2015), On computational modeling of visual saliency: examining what's right, and what's left, *Vision research*, *116*, 95–112.

Bruderlin, A., & Williams, L., (1995), Motion signal processing, *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 97–104, https://doi.org/10.1145/218380.218421

Bruneau, J., Olivier, A.-H., & Pettre, J., (2015), Going through, going around: a study on individual avoidance of groups, *IEEE transactions on visualization and computer graphics*, *21 4*, 520–528.

Bühler, M. A., & Lamontagne, A., (2018), Circumvention of pedestrians while walking in virtual and physical environments, *IEEE transactions on neural systems and rehabilitation engineering*, *26 9*, 1813–1822.

Burg, L., Lino, C., & Christie, M., (2020), Real-time anticipation of occlusions for automated camera control in toric space, *Computer Graphics Forum*, *39 2*, 523–533.

Burgoon, J. K., & Bacue, A. E., (2003), *Nonverbal communication skills.*, Lawrence Erlbaum Associates Publishers.

Buss, S. R., (2004), Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods, *IEEE Journal of Robotics and Automation*, *17 1-19*, 16.

Cañigueral, R., & Hamilton, A. F. d. C., (2019), The role of eye gaze during natural social interactions in typical and autistic people, *Frontiers in Psychology*, *10*, 560.

Cao, Z., Gao, H., Mangalam, K., Cai, Q.-Z., Vo, M., & Malik, J., (2020), Long-term human motion prediction with scene context, *European Conference on Computer Vision*, 387–404.

Casas, D., Tejera, M., Guillemaut, J.-Y., & Hilton, A., (2012), 4d parametric motion graphs for interactive animation, *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 103–110.

Chattaraj, U., Seyfried, A., & Chakroborty, P., (2009), Comparison of pedestrian fundamental diagram across cultures, *Advances in complex systems*, *12 03*, 393–405.

Chaumette, F., & Hutchinson, S., (2006), Visual servo control. i. basic approaches, *IEEE Robotics & Automation Magazine*, *13 4*, 82–90.

Chew, L. P., (1989), Constrained delaunay triangulations, *Algorithmica*, *4 1-4*, 97–108.

Chi, D., Costa, M., Zhao, L., & Badler, N., (2000), The emote model for effort and shape, *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 173–182.

Chinello, F., Malvezzi, M., Prattichizzo, D., & Pacchierotti, C., (2019), A modular wearable finger interface for cutaneous and kinesthetic interaction: control and evaluation, *IEEE Transactions on Industrial Electronics*, *67 1*, 706–716.

Chinello, F., Pacchierotti, C., Bimbo, J., Tsagarakis, N. G., & Prattichizzo, D., (2017a), Design and evaluation of a wearable skin stretch device for haptic guidance, *IEEE Robotics and Automation Letters*, *3 1*, 524–531.

Chinello, F., Pacchierotti, C., Malvezzi, M., & Prattichizzo, D., (2017b), A three revolute-revolute-spherical wearable fingertip cutaneous device for stiffness rendering, *IEEE Transactions on Haptics*, *11 1*, 39–50.

Chittaro, L., Ieronutti, L., Ranon, R., Siotto, E., & Visintini, D., (2010), A high-level tool for curators of 3d virtual visits and its application to a virtual exhibition of renaissance frescoes, *Proceedings of VAST*, *2010*, 11th.

Cho, S., Park, J., & Kwon, O., (2004), Gender differences in three dimensional gait analysis data from 98 healthy korean adults, *Clinical biomechanics*, *19 2*, 145–152.

Choi, B., i Ribera, R. B., Lewis, J. P., Seol, Y., Hong, S., Eom, H., Jung, S., & Noh, J., (2016), Sketchimo: sketch-based motion editing for articulated characters, *ACM Transactions on Graphics (TOG)*, *35 4*, 1–12.

Choi, K.-J., & Ko, H.-S., (2000), Online motion retargetting, *The Journal of Visualization and Computer Animation*, *11 5*, 223–235.

Chollet, M., & Scherer, S., (2017), Perception of virtual audiences, *IEEE computer graphics and applications*, *37 4*, 50–59.

Christie, M., & Olivier, P., (2006), Camera control in computer graphics: Models, techniques and applications, *EUROGRAPHICS*.

Christie, M., Olivier, P., & Normand, J.-M., (2008), Camera control in computer graphics, *Computer Graphics Forum*, *27 8*, 2197–2218.

Ciccone, L., Öztireli, C., & Sumner, R. W., (2019), Tangent-space optimization for interactive animation control, *ACM Transactions on Graphics (TOG)*, *38 4*, 1–10.

Cirio, G., Olivier, A.-H., Marchal, M., & Pettre, J., (2013), Kinematic evaluation of virtual walking trajectories, *IEEE transactions on visualization and computer graphics*, *19 4*, 671–680.

Clark, D. M., (1995), A cognitive model, *Social phobia: Diagnosis, assessment, and treatment*, 69–73.

Colombatto, C., van Buren, B., & Scholl, B. J., (2020), Gazing without eyes: a "stare-in-the-crowd" effect induced by simple geometric shapes, *Perception*, *49 7*, 782–792.

Colyer, S. L., Evans, M., Cosker, D. P., & Salo, A. I., (2018), A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system, *Sports medicine-open*, *4 1*, 1–15.

Cooper, R. M., Law, A. S., & Langton, S. R., (2013), Looking back at the stare-in-the-crowd effect: staring eyes do not capture attention in visual search, *Journal of vision*, *13 6*, 10–10.

Crehan, E. T., & Althoff, R. R., (2015), Measuring the stare-in-the-crowd effect: a new paradigm to study social perception, *Behavior research methods*, *47 4*, 994–1003.

Crehan, E. T., & Althoff, R. R., (2021), Me looking at you, looking at me: the stare-in-the-crowd effect and autism spectrum disorder, *Journal of Psychiatric Research*.

Croft, J. L., & Panchuk, D., (2018), Watch where you're going? interferer velocity and visual behavior predicts avoidance strategy during pedestrian encounters, *Journal of motor behavior*, *50 4*, 353–363.

Dael, N., Mortillaro, M., & Scherer, K. R., (2012), The body action and posture coding system (bap): development and reliability, *Journal of Nonverbal Behavior*, *36 2*, 97–121.

Debevec, P., (2006), Virtual cinematography: relighting through computation, *Computer*, *39 8*, 57–65.

de Gelder, B., De Borst, A., & Watson, R., (2015), The perception of emotion in body expressions, *Wiley Interdisciplinary Reviews: Cognitive Science*, *6 2*, 149–158.

Devigne, L., Aggravi, M., Bivaud, M., Balix, N., Teodorescu, S., Carlson, T., Spreters, T., Pacchierotti, C., & Babel, M., (2020), Power wheelchair navigation assistance using wearable vibrotactile haptics, *IEEE Transactions on Haptics*.

Doi, H., & Ueda, K., (2007), Searching for a perceived stare in the crowd, *Perception*, *36 5*, 773–780.

Dombre, E., & Khalil, W., (2013), *Robot manipulators: modeling, performance analysis and control*, John Wiley & Sons.

Durupinar, F., (2021), Perception of human motion similarity based on laban movement analysis, *ACM Symposium on Applied Perception 2021*, 1–7.

Durupinar, F., Kapadia, M., Deutsch, S., Neff, M., & Badler, N. I., (2016), Perform: perceptual approach for adding ocean personality to human motion using laban movement analysis, *ACM Transactions on Graphics (TOG)*, *36 1*, 1–16.

Duverné, T., Rougnant, T., Yondre, F. L., Berton, F., Bruneau, J., Zibrek, K., Pettré, J., Hoyet, L., & Olivier, A.-H., (2020), Effect of social settings on proxemics during social interactions in real and virtual conditions, *International Conference on Virtual Reality and Augmented Reality*, 3–19.

Ekman, P., & Friesen, W. V., (1978), *Facial action coding systems*, Consulting Psychologists Press.

Eom, H., Han, D., Shin, J. S., & Noh, J., (2019), Model predictive control with a visuomotor system for physics-based character animation, *ACM Transactions on Graphics (TOG)*, *39 1*, 1–11.

Espiau, B., Chaumette, F., & Rives, P., (1992), A new approach to visual servoing in robotics, *ieee Transactions on Robotics and Automation*, *8 3*, 313–326.

Faure, C., (2019), *Vers des environnements virtuels plus écologiques: étude des modifications du comportement moteur en réalité virtuelle lors de l'ajout d'informations haptiques par un mécanisme parallèle entraîné par câbles* [Doctoral dissertation, Université Laval, Canada], http://hdl.handle.net/20.500.11794/37886

Feng, A. W., Xu, Y., & Shapiro, A., (2012), An example-based motion synthesis technique for locomotion and object manipulation, *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 95–102.

Fetter, W. A., (1982), A progression of human figures simulated by computer graphics, *IEEE Computer Graphics and Applications*, *2 09*, 9–13.

Fischler, M. A., & Bolles, R. C., (1981), Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Graphics and Image Processing*, *10064*.

Fouhey, D. F., Scharstein, D., & Briggs, A. J., (2010), Multiple plane detection in image pairs using j-linkage, *2010 20th International Conference on Pattern Recognition*, 336–339.

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A., (2013), Leftward biases in picture scanning and line bisection: a gaze-contingent window study, *Vision Research*, *78*, 14–25, https://doi.org/https://doi.org/10.1016/j.visres.2012.12.001

Framorando, D., George, N., Kerzel, D., & Burra, N., (2016), Straight gaze facilitates face processing but does not cause involuntary attentional capture, *Visual cognition*, *24 7-8*, 381–391.

Frey, S., & Von Cranach, M., (1971), A method for measuring motor activity., *Zeitschrift für Experimentelle und Angewandte Psychologie*.

Galvane, Q., Christie, M., Lino, C., & Ronfard, R., (2015a), Camera-on-rails: automated computation of constrained camera paths, *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 151–157.

Galvane, Q., Lino, C., Christie, M., Fleureau, J., Servant, F., Tariolle, F. o.-l., & Guillotel, P., (2018), Directing cinematographic drones, *ACM Transactions on Graphics (TOG)*, *373*, 1–18.

Galvane, Q., Ronfard, R., Lino, C., & Christie, M., (2015b), Continuity editing for 3d animation, *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., & Sasse, M. A., (2003), The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment, *Proceedings of the SIGCHI conference on Human factors in computing systems*, 529–536.

Garcia, M., Ronfard, R., & Cani, M.-P., (2019), Spatial motion doodles: sketching animation in vr using hand gestures and laban motion analysis. *In Motion, interaction and games* (pp. 1–10).

Geijtenbeek, T., Van De Panne, M., & Van Der Stappen, A. F., (2013), Flexible muscle-based locomotion for bipedal creatures, *ACM Transactions on Graphics (TOG)*, *326*, 1–11.

Genay, A. C. S., Lécuyer, A., & Hachet, M., (2021), Being an avatar “for real”: a survey on virtual embodiment in augmented reality, *IEEE Transactions on Visualization and Computer Graphics*.

Gibson, J. J., (1958), Visually controlled locomotion and visual orientation in animals, *British journal of psychology*, *49 3*, 182–194.

Glardon, P., Boulic, R., & Thalmann, D., (2004), Pca-based walking engine using motion capture data, *Proceedings Computer Graphics International, 2004.*, 292–298, https://doi.org/10.1109/CGI.2004.1309224

Gleicher, M., (1998), Retargetting motion to new characters, *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, 33–42, https://doi.org/10.1145/280814.280820

Gleicher, M., (2001), Motion path editing, *Proceedings of the 2001 symposium on Interactive 3D graphics*, 195–202.

Gleicher, M., & Witkin, A., (1992), Through-the-lens camera control, *SIGGRAPH Comput. Graph.*, *26 2*, 331–340, https://doi.org/10.1145/142920.134088

Glémarec, Y., Lugrin, J.-L., Bosser, A.-G., Collins Jackson, A., Buche, C., & Latoschik, M. E., (2021), Indifferent or enthusiastic? virtual audiences animation and perception in virtual reality, *Frontiers in Virtual Reality*, *2*, 72.

Goldberg, L. R., (1990), An alternative" description of personality": the big-five factor structure., *Journal of personality and social psychology*, *59 6*, 1216.

Gonzalez-Franco, M., & Chou, P. A., (2014), Non-linear modeling of eye gaze perception as a function of gaze and head direction, *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 275–280.

Gonzalez-Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., Maselli, A., Banakou, D., Pelechano Gómez, N., Orts-Escolano, S., et al., (2020), The rocketbox library and the utility of freely available rigged avatars, *Frontiers in virtual reality*, *1 561558*, 1–23.

Guay, M., Ronfard, R., Gleicher, M., & Cani, M.-P., (2015), Space-time sketching of character animation, *ACM Transactions on Graphics (TOG)*, *34 4*, 1–10.

Guo, S., Southern, R., Chang, J., Greer, D., & Zhang, J. J., (2015), Adaptive motion synthesis for virtual characters: a survey, *The Visual Computer*, *31 5*, 497–512.

Hager, J. C., Ekman, P., & Friesen, W. V., (2002), Facial action coding system, *Salt Lake City, UT: A Human Face*.

Hall, E. T., (1969), *The hidden dimension*.

Halper, N., Helbing, R., & Strothotte, T., (2001), A camera engine for computer games: managing the trade-off between constraint satisfaction and frame coherence, *Computer Graphics Forum*, *20 3*, 174–183.

Hämäläinen, P., Eriksson, S., Tanskanen, E., Kyrki, V., & Lehtinen, J., (2014), Online motion synthesis using sequential monte carlo, *ACM Transactions on Graphics (TOG)*, *33 4*, 1–12.

Harrigan, J., Rosenthal, R., Scherer, K. R., & Scherer, K., (2008), *New handbook of methods in nonverbal behavior research*, Oxford University Press.

Heck, R., & Gleicher, M., (2007), Parametric motion graphs, *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, 129–136.

Hessels, R. S., van Doorn, A. J., Benjamins, J. S., Holleman, G. A., & Hooge, I. T., (2020), Task-related gaze control in human crowd navigation, *Attention, Perception, & Psychophysics*, 1–20.

Hickey, L. F., Springer, W. E., & Cundari, F. L., (1968), *A development in cockpit geometry evaluation* (tech. rep.), BOEING CO SEATTLE WA.

Hinde, R. A., (1972), *Non-verbal communication*, Cambridge University Press.

Ho, E. S. L., Komura, T., & Tai, C.-L., (2010), Spatial relationship preserving character motion adaptation, *ACM Trans. Graph.*, *29 4*, https://doi.org/10.1145/1778765.1778770

Hodge, E. M., Tabrizi, M., Farwell, M. A., & Wuensch, K. L., (2008), Virtual reality classrooms: strategies for creating a social presence, *International Journal of Social Sciences*, *2 2*, 105–109.

Hoffmann, L., Krämer, N. C., Lam-Chi, A., & Kopp, S., (2009), Media equation revisited: do users show polite reactions towards an embodied agent?, *Intelligent Virtual Agents: 9th International Conference, IVA 2009 Amsterdam, The Netherlands, September 14-16, 2009 Proceedings 9*, 159–165.

Holden, D., Habibie, I., Kusajima, I., & Komura, T., (2017a), Fast neural style transfer for motion data, *IEEE computer graphics and applications*, *37 4*, 42–49.

Holden, D., Kanoun, O., Perepichka, M., & Popa, T., (2020), Learned motion matching, *ACM Transactions on Graphics (TOG)*, *39 4*, 53–1.

Holden, D., Komura, T., & Saito, J., (2017b), Phase-functioned neural networks for character control, *ACM Transactions on Graphics (TOG)*, *36 4*, 1–13.

Holden, D., Saito, J., & Komura, T., (2016a), A deep learning framework for character motion synthesis and editing, *ACM Transactions on Graphics (TOG)*, *35 4*, 1–11.

Holden, D., Saito, J., & Komura, T., (2016b), Learning inverse rig mappings by nonlinear regression, *IEEE transactions on visualization and computer graphics*, *23 3*, 1167–1178.

Huang, H., Lischinski, D., Hao, Z., Gong, M., Christie, M., & Cohen-Or, D., (2016), Trip synopsis: 60km in 60sec, *Computer Graphics Forum*, *35*, 107–116.

Huang, P.-H., & Wong, S.-K., (2018), Emotional virtual crowd on task completion in virtual markets, *Computer Animation and Virtual Worlds*, *29 3-4*, e1818.

Huang, Y., & Kallmann, M., (2015), Planning motions and placements for virtual demonstrators, *IEEE transactions on visualization and computer graphics*, *22 5*, 1568–1579.

Huber, M., Su, Y.-H., Krüger, M., Faschian, K., Glasauer, S., & Hermsdörfer, J., (2014), Adjustments of speed and path when avoiding collisions with another pedestrian, *PloS one*, *9 2*.

Iachini, T., Coello, Y., Frassinetti, F., Senese, V. P., Galante, F., & Ruggiero, G., (2016a), Peripersonal and interpersonal space in virtual and real environments: effects of gender and age, *Journal of Environmental Psychology*, *45*, 154–164.

Iachini, T., Coello, Y., Frassinetti, F., Senese, V. P., Galante, F., & Ruggiero, G., (2016b), Peripersonal and interpersonal space in virtual and real environments: effects of gender and age, *Journal of Environmental Psychology*, *45*, 154–164.

Jiang, H., Christie, M., Wang, X., Liu, L., Wang, B., & Chen, B., (2021), Camera keyframing with style and control, *ACM Transactions on Graphics (TOG)*, *40 6*, 1–13.

Jiang, H., Wang, B., Wang, X., Christie, M., & Chen, B., (2020), Example-driven virtual cinematography by learning camera behaviors, *ACM Transactions on Graphics (TOG)*, *39 3*.

Kahlon, S., Lindner, P., & Nordgreen, T., (2019), Virtual reality exposure therapy for adolescents with fear of public speaking: a non-randomized feasibility and pilot study, *Child and adolescent psychiatry and mental health*, *13 1*, 47.

Keogh, E., Palpanas, T., Zordan, V. B., Gunopulos, D., & Cardle, M., (2004), Indexing large human-motion databases, *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 780–791.

Kilteni, K., Groten, R., & Slater, M., (2012), The sense of embodiment in virtual reality, *Presence: Teleoperators and Virtual Environments*, *21 4*, 373–387.

Kim, M., Hyun, K., Kim, J., & Lee, J., (2009), Synchronized multi-character motion editing, *ACM transactions on graphics (TOG)*, *28 3*, 1–9.

Kimura, A., Yonetani, R., & Hirayama, T., (2013), Computational models of human visual attention and their implementations: a survey, *IEICE TRANSACTIONS on Information and Systems*, *96 3*, 562–578.

Kleinke, C. L., (1986), Gaze and eye contact: a research review., *Psychological bulletin*, *100 1*, 78.

Knoblich, G., & Sebanz, N., (2008), Evolving intentions for social interaction: from entrainment to joint action, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363 1499*, 2021–2031.

Knorr, A. G., Willacker, L., Hermsdörfer, J., Glasauer, S., & Krüger, M., (2016), Influence of person-and situation-specific characteristics on collision avoidance behavior in human locomotion., *Journal of experimental psychology: human perception and performance*, *42 9*, 1332.

Kovar, L., & Gleicher, M., (2003), Flexible automatic motion blending with registration curves, *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 214–224.

Kovar, L., & Gleicher, M., (2004), Automated extraction and parameterization of motions in large data sets, *ACM Transactions on Graphics (ToG)*, *23 3*, 559–568.

Kovar, L., Gleicher, M., & Pighin, F., (2002), Motion graphs, *ACM Trans. Graph.*, *21 3*, 473–482, https://doi.org/10.1145/566654.566605

Krogmeier, C., Mousas, C., & Whittinghill, D., (2019), Human, virtual human, bump! a preliminary study on haptic feedback, *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 1032–1033.

Krum, D. M., Kang, S.-H., & Phan, T., (2018), Influences on the elicitation of interpersonal space with virtual humans, *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 223–9.

Kwiatkowski, A., Alvarado, E., Kalogeiton, V., Liu, C. K., Pettré, J., van de Panne, M., & Cani, M.-P., (2022), A survey on reinforcement learning methods in character animation, *arXiv preprint arXiv:2203.04735*.

Laban, R., & Ullmann, L., (1971), The mastery of movement.

Lamarche, F., (2009), Topoplan: a topological path planner for real time human navigation under floor and ceiling constraints, *Computer Graphics Forum*, *28 2*, 649–658.

Lange, B., & Pauli, P., (2019), Social anxiety changes the way we move—a social approach-avoidance task in a virtual reality cave system, *PloS One*, *14 12*, e0226805.

Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., & Pollard, N. S., (2002), Interactive control of avatars animated with human motion data, *ACM Trans. Graph.*, *21 3*, 491–500, https://doi.org/10.1145/566654.566607

Lee, Y., Wampler, K., Bernstein, G., Popović, J., & Popović, Z., (2010), Motion fields for interactive character locomotion, *ACM Trans. Graph.*, *29 6*, https://doi.org/10.1145/1882261.1866160

Li, J., Kong, Y., Röggla, T., De Simone, F., Ananthanarayan, S., De Ridder, H., El Ali, A., & Cesar, P., (2019), Measuring and understanding photo sharing experiences in social virtual reality, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Li, J., Yang, T., & Yu, J., (2017), Random sampling and model competition for guaranteed multiple consensus sets estimation, *International Journal of Advanced Robotic Systems*, *14 1*.

Li, T.-Y., & Cheng, C.-C., (2008), Real-time camera planning for navigation in virtual environments, *International Symposium on Smart Graphics*, 118–129.

Liebowitz, M. R., (1987), Social phobia., *Modern problems of pharmacopsychiatry*.

Lindeman, R. W., Page, R., Yanagida, Y., & Sibert, J. L., (2004), Towards full-body haptic feedback: the design and deployment of a spatialized vibrotactile feedback system, *Proceedings of the ACM symposium on Virtual reality software and technology*, 146–149.

Lindeman, R. W., Yanagida, Y., Noma, H., & Hosaka, K., (2006), Wearable vibrotactile systems for virtual contact and information display, *Virtual Reality*, *9 2-3*, 203–213.

Lino, C., Christie, M., Lamarche, F., Schofield, G., & Olivier, P., (2010), A real-time cinematography system for interactive 3d environments, *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 139–148.

Liu, C. K., Hertzmann, A., & Popović, Z., (2005), Learning physics-based motion style with nonlinear inverse optimization, *ACM Transactions on Graphics (TOG)*, *24 3*, 1071–1081.

Louarn, A., Christie, M., & Lamarche, F., (2018), Automated staging for virtual cinematography, *Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games*, 4.

Louison, C., Ferlay, F., & Mestre, D. R., (2018), Spatialized vibrotactile feedback improves goal-directed movements in cluttered virtual environments, *International Journal of Human–Computer Interaction*, *34 11*, 1015–1031.

Lynch, S. D., Kulpa, R., Meerhoff, L. A., Pettre, J., Cretual, A., & Olivier, A.-H., (2017), Collision avoidance behavior between walkers: global and local motion cues, *IEEE transactions on visualization and computer graphics*, *24 7*, 2078–2088.

Manor, B. R., & Gordon, E., (2003), Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks, *Journal of Neuroscience Methods*, *128 1*, 85–93, https://www.sciencedirect.com/science/article/pii/S0165027003001511

Mar, R. A., (2011), The neural bases of social cognition and story comprehension, *Annual review of psychology*, *62*, 103–134.

Marschner, L., Pannasch, S., Schulz, J., & Graupner, S.-T., (2015), Social communication with virtual agents: the effects of body and gaze direction on attention and emotional responding in human observers, *International Journal of Psychophysiology*, *97 2*, 85–92.

McDonnell, R., Jörg, S., Power, J., Newell, F., & O'Sullivan, C., (2008), Evaluating the emotional content of human motions on real and virtual characters, 67–74, https://doi.org/10.1145/1394281.1394294

Mehrabian, A., (1971), Nonverbal communication., *Nebraska symposium on motivation*.

Men, Q., Shum, H. P., Ho, E. S., & Leung, H., (2022), Gan-based reactive motion synthesis with class-aware discriminators for human–human interaction, *Computers & Graphics*, *102*, 634–645.

Ménardais, S., Multon, F., Kulpa, R., & Arnaldi, B., (2004), Motion blending for real-time animation while accounting for the environment, *Proceedings Computer Graphics International, 2004.*, 156–159.

Mestre, D. R., Louison, C., & Ferlay, F., (2016), The contribution of a virtual self and vibrotactile feedback to walking through virtual apertures, *International Conference on Human-Computer Interaction*, 222–232.

Moeslund, T. B., Hilton, A., & Krüger, V., (2006), A survey of advances in vision-based human motion capture and analysis, *Computer vision and image understanding*, *104 2-3*, 90–126.

Mohammad, Y., & Nishida, T., (2008), Reactive gaze control for natural human-robot interactions, *2008 IEEE Conference on Robotics, Automation and Mechatronics*, 47–54.

Mordatch, I., Todorov, E., & Popović, Z., (2012), Discovery of complex behaviors through contact-invariant optimization, *ACM Transactions on Graphics (TOG)*, *31 4*, 1–8.

Mori, M., (1970), Bukimi no tani [the uncanny valley], *Energy*, *7*, 33–35.

Mourot, L., Hoyet, L., Clerc, F. L., Schnitzler, F., & Hellier, P., (2021), A survey on deep learning for skeleton-based human animation, *arXiv preprint arXiv:2110.06901*.

Naderi, K., Rajamäki, J., & Hämäläinen, P., (2017), Discovering and synthesizing humanoid climbing movements, *ACM Transactions on Graphics (TOG)*, *36 4*, 1–11.

Narang, S., Best, A., Randhavane, T., Shapiro, A., & Manocha, D., (2016), Pedvr: simulating gaze-based interactions between a real user and virtual crowds, *Proceedings of the 22nd ACM conference on virtual reality software and technology*, 91–100.

Neff, M., Wang, Y., Abbott, R., & Walker, M., (2010), Evaluating the effect of gesture and language on personality perception in conversational agents, *In* J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents* (pp. 222–235), Springer Berlin Heidelberg.

Nieuwenhuisen, D., & Overmars, M. H., (2004), Motion planning for camera movements, *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, *4*, 3870–3876.

Nummenmaa, L., Hyönä, J., & Hietanen, J. K., (2009), I'll walk this way: eyes reveal the direction of locomotion and make passersby look and go the other way, *Psychological science*, *20 12*, 1454–1458.

Oh, C. S., Bailenson, J. N., & Welch, G. F., (2018), A systematic review of social presence: definition, antecedents, and implications, *Frontiers in Robotics and AI*, *5*, 114.

Oliva, R., & Pelechano, N., (2011), Automatic generation of suboptimal navmeshes, *International Conference on Motion in Games*, 328–339.

Olivier, A.-H., Bruneau, J., Kulpa, R., & Pettré, J., (2017), Walking with virtual people: evaluation of locomotion interfaces in dynamic environments, *IEEE transactions on visualization and computer graphics*, *24 7*, 2251–2263.

Olivier, A.-H., Marin, A., Crétual, A., Berthoz, A., & Pettré, J., (2013), Collision avoidance between two walkers: role-dependent strategies, *Gait & posture*, *38 4*, 751–756.

Olivier, A.-H., Marin, A., Crétual, A., & Pettré, J., (2012), Minimal predicted distance: a common metric for collision avoidance during pairwise interactions between walkers, *Gait & posture*, *36 3*, 399–404.

O'Rourke, J., (1987), *Art gallery theorems and algorithms*, Oxford University Press, Inc.

Oskam, T., Sumner, R. W., Thuerey, N., & Gross, M., (2009), Visibility transition planning for dynamic camera control, *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 55–65.

Ossandón, J. P., Onat, S., & König, P., (2014), Spatial biases in viewing behavior, *Journal of Vision*, *14 2*, 20–20, https://doi.org/10.1167/14.2.20

Pacchierotti, C., Sinclair, S., Solazzi, M., Frisoli, A., Hayward, V., & Prattichizzo, D., (2017), Wearable haptic systems for the fingertip and the hand: taxonomy, review, and perspectives, *IEEE Transactions on Haptics*, *10 4*, 580–600.

Palanica, A., & Itier, R., (2011a), Measuring the stare-in-the-crowd effect using eye-tracking: effects of task demands, *Journal of Vision*, *11 11*, 1327–1327.

Palanica, A., & Itier, R. J., (2011b), Searching for a perceived gaze direction using eye tracking, *Journal of Vision*, *11 2*, 19–19.

Pan, X., & Hamilton, A. F. d. C., (2018), Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape, *British Journal of Psychology*, *109 3*, 395–417.

Park, S. I., Shin, H. J., Kim, T. H., & Shin, S. Y., (2004), On-line motion blending for real-time locomotion generation, *Computer Animation and Virtual Worlds*, *15 3-4*, 125–138.

Patla, A. E., (1997), Understanding the roles of vision in the control of human locomotion, *Gait & Posture*, *5 1*, 54–69.

Peng, Z., Genewein, T., & Braun, D. A., (2014), Assessing randomness and complexity in human motion trajectories through analysis of symbolic sequences, *Frontiers in human neuroscience*, *8*, 168.

Perrinet, J., Olivier, A.-H., & Pettre, J., (2013), Walk with me: interactions in emotional walking situations, a pilot study, *Proceedings - SAP 2013: ACM Symposium on Applied Perception*, https://doi.org/10.1145/2492494.2492507

Pittig, A., Arch, J. J., Lam, C. W., & Craske, M. G., (2013), Heart rate and heart rate variability in panic, social anxiety, obsessive–compulsive, and generalized anxiety disorders at baseline and in response to relaxation and hyperventilation, *International journal of psychophysiology*, *87 1*, 19–27.

Raimbaud, P., Jovane, A., Zibrek, K., Pacchierotti, C., Christie, M., Hoyet, L., Pettré, J., & Olivier, A.-H., (2021), Reactive virtual agents: a viewpoint-driven approach for bodily nonverbal communication, *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 164–166.

Raisamo, R., Salminen, K., Rantala, J., Farooq, A., & Ziat, M., (2022), Interpersonal haptic communication: review and directions for the future, *International Journal*

of Human-Computer Studies, *166*, 102881, https://doi.org/https://doi.org/10.1016/j.ijhcs.2022.102881

Ramamoorthy, N., Plaisted-Grant, K., & Davis, G., (2019), Fractionating the stare-in-the-crowd effect: two distinct, obligatory biases in search for gaze., *Journal of Experimental Psychology: Human Perception and Performance*, *458*, 1015.

Randhavane, T., Bera, A., Kapsaskis, K., Sheth, R., Gray, K., & Manocha, D., (2019a), Eva: generating emotional behavior of virtual agents using expressive features of gait and gaze, *ACM Symposium on Applied Perception 2019*, https://doi.org/10.1145/3343036.3343129

Randhavane, T., Bhattacharya, U., Kapsaskis, K., Gray, K., Bera, A., & Manocha, D., (2019b), Identifying emotions from walking using affective and deep features, *arXiv preprint arXiv:1906.11884*.

Rio, K., & Warren, W. H., (2014), The visual coupling between neighbors in real and virtual crowds, *Transportation Research Procedia*, *2*, 132–140.

Rio, K. W., Dachner, G. C., & Warren, W. H., (2018), Local interactions underlying collective motion in human crowds, *Proceedings of the Royal Society B: Biological Sciences*, *2851878*, 20180611.

Roether, C. L., Omlor, L., Christensen, A., & Giese, M. A., (2009), Critical features for the perception of emotion from gait, *Journal of vision*, *96*, 15–15.

Rose, C., Cohen, M. F., & Bodenheimer, B., (1998), Verbs and adverbs: multidimensional motion interpolation, *IEEE Computer Graphics and Applications*, *185*, 32–40.

Rose III, C. F., Sloan, P.-P. J., & Cohen, M. F., (2001), Artist-directed inverse-kinematics using radial basis function interpolation, *Computer graphics forum*, *203*, 239–250.

Roth, D., & Latoschik, M. E., (2020), Construction of the virtual embodiment questionnaire (veq), *IEEE Transactions on Visualization and Computer Graphics*, *2612*, 3546–3556, https://doi.org/10.1109/TVCG.2020.3023603

Roth, D., Kullmann, P., Bente, G., Gall, D., & Latoschik, M. E., (2018), Effects of hybrid and synthetic social gaze in avatar-mediated interactions, *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 103–108.

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., Mutlu, B., & McDonnell, R., (2015), A review of eye gaze in virtual agents, social robotics and hci: behaviour generation, user interaction and perception, *Computer Graphics Forum*, *346*, 299–326, https://doi.org/https://doi.org/10.1111/cgf.12603

Safonova, A., & Hodgins, J. K., (2007), Construction and optimal search of interpolated motion graphs. *In Acm siggraph 2007 papers* (106–es).

Salazar, S. V., Pacchierotti, C., de Tinguy, X., Maciel, A., & Marchal, M., (2020), Altering the stiffness, friction, and shape perception of tangible objects in virtual reality using wearable haptics, *IEEE transactions on haptics*, *13 1*, 167–174.

Scheggi, S., Aggravi, M., & Prattichizzo, D., (2016), Cooperative navigation for mixed human–robot teams using haptic feedback, *IEEE Transactions on Human-Machine Systems*, *47 4*, 462–473.

Schorr, S. B., & Okamura, A. M., (2017), Three-dimensional skin deformation as force substitution: wearable device design and performance during haptic exploration of virtual environments, *IEEE transactions on haptics*, *10 3*, 418–430.

Schuetzler, R. M., Grimes, G. M., & Giboney, J. S., (2018), An investigation of conversational agent relevance, presence, and engagement (N. O. A. for Information Systems, Ed.), *Americas conference on information systems 2018 proceedings*.

Schulze, L., Lobmaier, J. S., Arnold, M., & Renneberg, B., (2013), All eyes on me?! social anxiety and self-directed perception of eye gaze [PMID: 23438447], *Cognition and Emotion*, *27 7*, 1305–1313, https://doi.org/10.1080/02699931.2013.773881

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M., (2021), Voice in human–agent interaction: a survey, *ACM Computing Surveys (CSUR)*, *54 4*, 1–43.

Seyfried, A., Steffen, B., Klingsch, W., & Boltes, M., (2005), The fundamental diagram of pedestrian movement revisited, *Journal of Statistical Mechanics: Theory and Experiment*, *2005 10*, P10002.

Shao, W., & Ng-Thow-Hing, V., (2003), A general joint component framework for realistic articulation in human characters, *Proceedings of the 2003 symposium on Interactive 3D graphics*, 11–18.

Shapiro, A., Cao, Y., & Faloutsos, P., (2006), Style components., *Graphics interface*, *2006*, 33–39.

Sharma, S., Verma, S., Kumar, M., & Sharma, L., (2019), Use of motion capture in 3d animation: motion capture systems, challenges, and recent trends, *2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)*, 289–294.

Shum, H. P., Komura, T., Shiraishi, M., & Yamazaki, S., (2008), Interaction patches for multi-character animation, *ACM Transactions on Graphics (TOG)*, *27 5*, 1–8.

Silva, W. S., Aravind, G., Sangani, S., & Lamontagne, A., (2018), Healthy young adults implement distinctive avoidance strategies while walking and circumventing virtual human vs. non-human obstacles in a virtual environment, *Gait & posture*, *61*, 294–300.

Silver, C. A., Tatler, B. W., Chakravarthi, R., & Timmermans, B., (2020), Social agency as a continuum, *Psychonomic Bulletin & Review*, 1–20.

Slater, M., (2003), A note on presence terminology, *Presence connect*, *3 3*, 1–5.

Slater, M., & Usoh, M., (1993), Presence in immersive virtual environments, *Proceedings of IEEE virtual reality annual international symposium*, 90–96.

Slater, M., Usoh, M., & Steed, A., (1994), Depth of presence in virtual environments, *Presence: Teleoperators & Virtual Environments*, *3 2*, 130–144.

Smith, H. J., & Neff, M., (2017), Understanding the impact of animated gesture performance on personality perceptions, *ACM Transactions on Graphics (TOG)*, *36 4*, 1–12.

Sørensen, T. J., (1948), A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons, *Biol. Skr.*, *5*, 1–34.

Starke, S., Zhao, Y., Komura, T., & Zaman, K., (2020), Local motion phases for learning multi-contact character movements, *ACM Transactions on Graphics (TOG)*, *39 4*, 54–1.

Starke, S., Zhao, Y., Zinno, F., & Komura, T., (2021), Neural animation layering for synthesizing martial arts movements, *ACM Transactions on Graphics (TOG)*, *40 4*, 1–16.

Sun, Z., Yu, W., Zhou, J., & Shen, M., (2017), Perceiving crowd attention: gaze following in human crowds with conflicting cues, *Attention, Perception, & Psychophysics*, *79 4*, 1039–1049.

Switonski, A., Josinski, H., & Wojciechowski, K., (2019), Dynamic time warping in classification and selection of motion capture data, *Multidimensional Systems and Signal Processing*, *30 3*, 1437–1468.

Tagliasacchi, A., Alhashim, I., Olson, M., & Zhang, H., (2012), Mean curvature skeletons, *Computer Graphics Forum*, *31 5*, 1735–1744.

Teyssier, M., Bailly, G., Pelachaud, C., & Lecolinet, E., (2020), Conveying emotions through device-initiated touch, *IEEE Transactions on Affective Computing*.

Thompson, R., & Bowen, C. J., (2009), *Grammar of the edit* (Vol. 13), Taylor & Francis.

Unuma, M., Anjyo, K., & Takeuchi, R., (1995), Fourier principles for emotion-based human figure animation, *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 91–96, https://doi.org/10.1145/218380.218419

Unzueta, L., Peinado, M., Boulic, R., & Suescun, Á., (2008), Full-body performance animation with sequential inverse kinematics, *Graphical models*, *70 5*, 87–104.

Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., & Fua, P., (2004), Style-based motion synthesis, *Computer Graphics Forum*, *23 4*, 799–812.

Usoh, M., Catena, E., Arman, S., & Slater, M., (2000), Using presence questionnaires in reality, *Presence: Teleoperators & Virtual Environments*, *9 5*, 497–503.

van den Hengel, A., Hill, R., Ward, B., Cichowski, A., Detmold, H., Madden, C., Dick, A., & Bastian, J., (2009), Automatic camera placement for large scale surveillance networks, *2009 Workshop on Applications of Computer Vision (WACV)*, 1–6.

Van der Kruk, E., & Reijne, M. M., (2018), Accuracy of human motion capture systems for sport applications; state-of-the-art review, *European journal of sport science*, *18 6*, 806–819.

Van Welbergen, H., Van Basten, B. J. H., Egges, A., Ruttkay, Z. M., & Overmars, M. H., (2010), Real time animation of virtual humans: a trade-off between naturalness and control, *Computer Graphics Forum*, *29 8*, 2530–2554, https://doi.org/https://doi.org/10.1111/j.1467-8659.2010.01822.x

Villegas, R., Yang, J., Ceylan, D., & Lee, H., (2018), Neural kinematic networks for unsupervised motion retargetting, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8639–8648, https://doi.org/10.1109/CVPR.2018.00901

Volonte, M., Hsu, Y. C., Liu, K.-y., Mazer, J., Wong, S.-K., & Babu, S., (2020), Effects of interacting with a crowd of emotional virtual humans on users' affective and non-verbal behaviors, *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*.

Von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S.-H., (2010), "it doesn't matter what you are!" explaining social effects of agents and avatars., *Computers in Human Behavior*.

Von Grünau, M., & Anston, C., (1995), The detection of gaze direction: a stare-in-the-crowd effect, *Perception*, *24 11*, 1297–1313.

Wang, X., Chen, Q., & Wang, W., (2014), 3d human motion editing and synthesis: a survey, *Computational and Mathematical methods in medicine*, *2014*.

Warren, W., (1998), Visually controlled locomotion: 40 years later, *Ecological Psychology*, *10 3-4*, 177–219.

Westheimer, G., (1954), Eye movement responses to a horizontally moving visual stimulus, *AMA Archives of Ophthalmology*, *52 6*, 932–941.

Wieser, M. J., Pauli, P., Grosseibl, M., Molzow, I., & Mühlberger, A., (2010), Virtual social interactions in social anxiety—the impact of sex, gaze, and interpersonal distance, *Cyberpsychology, Behavior, and Social Networking*, *13 5*, 547–554.

Wiley, D. J., & Hahn, J. K., (1997), Interpolation synthesis of articulated figure motion, *IEEE Computer Graphics and Applications*, *17 6*, 39–45.

Wilmut, K., & Barnett, A. L., (2010), Locomotor adjustments when navigating through apertures, *Human Movement Science*, *29 2*, 289–298.

Witkin, A., & Popovic, Z., (1995), Motion warping, *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 105–108, https://doi.org/10.1145/218380.218422

Xia, S., Wang, C., Chai, J., & Hodgins, J., (2015), Realtime style transfer for unlabeled heterogeneous human motion, *ACM Transactions on Graphics (TOG)*, *34 4*, 1–10.

Xiang Xu, K. Z., Min Huang, (2011), Automatic generated navigation mesh algorithm on 3d game scene, *Procedia Engineering*, *15*, 3215–3219.

Yumer, M. E., & Mitra, N. J., (2016), Spectral style transfer for human motion between independent actions, *ACM Transactions on Graphics (TOG)*, *35 4*, 1–8.

Zhou, F., & De la Torre, F., (2012), Generalized time warping for multi-modal alignment of human motion, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1282–1289, https://doi.org/10.1109/CVPR.2012.6247812

Zhou, H., & Hu, H., (2008), Human motion tracking for rehabilitation—a survey, *Biomedical signal processing and control*, *3 1*, 1–18.

Zibrek, K., Niay, B., Olivier, A.-H., Hoyet, L., Pettre, J., & McDonnell, R., (2020), The effect of gender and attractiveness of motion on proximity in virtual reality, *ACM Transactions on Applied Perception (TAP)*, *17 4*, 1–15.

Appendix

A Virtual Character Gaze in Virtual Reality: Exploring the Stare-in-the-crowd Effect

In this Appendix we present and discuss the detailed result, for gaze behaviors that were not addressed in the main Chapter 3.

A.1 Results and discussion

As we have already introduced in Chapter 3, we study five comparisons on the collected data, related to three categories: (1) the stare-in-the-crowd effect in static conditions, (2) *catching someone else staring* and (3) *being caught staring* phenomena, in line with Crehan et al. [2015].

- For (1), we compared the averted to the directed gaze conditions – A vs. D. Then, we compared each static condition with each dynamic.
- For (2), averted versus averted-then-directed – A vs. AD, and directed versus averted-then-directed – D vs. AD.
- For (3), the averted versus directed-then-averted – A vs. DA, and directed versus directed-then-averted – D vs. DA.

In this Appendix we complete the discussion with the three remaining cases: A vs. AD, D vs. AD and D vs. DA.

As a reminder, for pairwise comparisons, we ran dependent t-tests for paired samples on the six metrics we described in Section 3.2.6 as continuous variables. Such tests guarantee conservative results in the comparison between different gaze conditions. The normal distribution assumption was verified for 25 of our 30 dependent paired samples when running a Shapiro-Wilk test: we ran Student's t-tests for these samples, and Wilcoxon signed rank tests for the remaining ones. Due to our multiple comparison design, we con-

Table A.1 – Gaze metric results - comparison of D vs. DA conditions

	Directed		Directed-then-Averted			
Metric	Mean (SD)		Mean (SD)	p-value	t	η_p^2
Dwell time	1570 (864)	↘	808 (363)	<0.00001 ***	6.83	0.62
Fixation count	2.35 (1.03)	↘	1.56 (0.56)	<0.00001 ***	6.17	0.57
1st fix. duration	552 (185)		483 (165)	0.07638	1.84	0.10
1st fix. time	4969 (1402)		4847 (1307)	0.69058	0.40	0.01
2nd fix. duration	554 (214)	↘	374 (95)	0.00016 **	4.32	0.39
2nd fix. time	6861 (1785)		7773 (1724)	0.05246	-2.03	0.12

Time and duration in ms.

ducted a Bonferroni correction which changed our target significance level from α=0.05 to α=0.00166. The obtained values are shown by comparison of pairs in Table A.2 for A vs. AD, Table A.3 for D vs. AD to A.3 and Table A.1 for D vs. DA. We repeat that, for each metric, they contain the means and standard deviations, along with significance level, plus statistics and effect size (both when doing Student's t-test). Results are based on the averages obtained by each user *across all trials that share the same gazing conditions regardless of position*, *i.e.*, 18 in total for each condition. In these tables, a symbol * indicates a p-value <0.00166, ** a p-value <0.00033, and *** a p-value <0.00003.

Comparison D vs. DA: interpretation. As shown in Table A.1, *dwell time, fixation count and second fixation duration* were significantly different between directed and directed-then-averted conditions, with lower values in the latter. **These results confirm the stare-in-the-crowd effect**: indeed, in a directed-then-averted condition, once the first fixation on the active agent had started, its gaze remained averted, thus significant differences are consistent with the ones observed for these metrics on the averted vs. directed comparison. In a similar way, in both conditions, the agent's gaze was directed before the first fixation started, which can explain the absence of the significant difference for the *first fixation time*. After that, for the *first fixation duration* and the *second fixation time*, the absence of significant difference between these two conditions is consistent with the interpretation given for averted vs. directed-then-averted conditions and could thus be explained the following: the gaze change of the active agent that occurred at the beginning of the first fixation could have captured the virtual reality users' attention at a level not significantly different to the one caused by a directed gaze for the first fixation in

Table A.2 – Gaze metrics results - comparison of A vs. AD conditions

	Averted		Averted-then-Directed			
Metric	Mean (SD)		Mean (SD)	p-value	t	η_p^2
Dwell time	504 (175)	↗	1503 (789)	<0.00001 ***	wilc.	wilc.
Fixation count	1.15 (0.29)	↗	2.18 (0.79)	<0.00001 ***	-7.24	0.64
1st fix. duration	332 (77)	↗	544 (176)	<0.00001 ***	-6.22	0.57
1st fix. time	5173 (1213)		5371 (1229)	0.87121	wilc.	wilc.
2nd fix. duration	407 (282)	↗	644 (226)	0.00011 **	wilc.	wilc.
2nd fix. time	8602 (1395)	↘	6978 (1544)	0.00032 **	4.08	0.36

Time and duration in ms.

terms of duration, and could have nonetheless made them check back towards this agent as soon as in the directed gaze condition, therefore through an early second fixation on it.

In addition, compared to our results, Crehan et al. [2015] did not observe the stare-in-the-crowd effect in all the metrics, since they found no effect of dwell time or fixation count. However, as we did, they found a significant difference for the second fixation duration. Our differences may come from the specifics of our setup, *e.g.* using virtual reality that adds depth and space information, unlike photographs.

Comparison A vs. AD: interpretation. As shown in Table A.2, *dwell time, fixation count, first fixation duration, second fixation duration and second fixation time* were significantly different between conditions averted and averted-then-directed, with lower value for the second fixation time and higher values for the other metrics in the averted-then-directed condition. **These results confirm the stare-in-the-crowd effect in virtual reality**: indeed in an averted-then-directed condition, once started the first fixation on the active agent, its gaze remains a directed gaze, therefore significant differences are consistent with the ones observed for these metrics on averted vs. directed comparison. Moreover, *first fixation time* was not significantly different between the two conditions, which is coherent since in both conditions the active virtual agent starts with an averted gaze.

In addition, we found similar results as with Crehan et al.'s ones [2015], except from the fact that they did not observe a higher level of first fixation duration in the dynamic

Table A.3 – Gaze metrics results - comparison of D vs. AD conditions

	Directed	Averted-then-Directed			
Metric	Mean (SD)	Mean (SD)	p-value	t	η_p^2
Dwell time	1570 (864)	1503 (789)	0.45729	0.75	0.02
Fixation count	2.35 (1.03)	2.18 (0.79)	0.15961	1.44	0.07
1st fix. duration	552 (185)	544 (176)	0.85695	0.18	0.00
1st fix. time	4969 (1402)	5371 (1229)	0.16703	-1.41	0.06
2nd fix. duration	554 (214)	644 (226)	0.13021	-1.56	0.08
2nd fix. time	6861 (1785)	6978 (1544)	0.75169	0.32	0.00

Time and duration in ms.

condition. Similarly, our differences may come from the specifics of our setup, *e.g.* using virtual reality that adds depth and space information, unlike photographs.

Comparison D vs. AD: interpretation. As shown in Table A.3, *dwell time, fixation count, first fixation duration, second fixation duration and second fixation time* were not significantly different between the two conditions. These results are coherent since in the averted-then-directed condition, once the first fixation on the active agent had started, its gaze remained directed, *i.e.*, with a gaze similar to the directed condition. Finally, for the *first fixation time* metric, there was no significant difference; which is discussed in the next paragraph. In addition, all our results are coherent with Crehan et al.'s ones [2015].

Active agent's position effect. In addition to these results that average all the data by gaze condition, our metrics can also be computed based on the averages obtained by each user across the trials that share both the same viewing conditions and the same position of the "active agent" in the crowd and therefore in the user's field of view – two repetitions in total for each condition. Due to the variability of the number of fixations across conditions and users, dwell time and fixation count metrics were preferred here over fixation time and duration metrics, since the former ones can always be computed even when no fixations occurred on the expected agent during the trials – in that case, missing values would be reported for the other metrics when computing averages. For these nine

Table A.4 – Metrics comparison for each position A vs. D - dwell time and fixation count

	Dwell time metric	Fixation count metric
Paired samples for t-test	p-value	p-value
Left-close A - Left-close D	0.00203 *	0.00078 **
Left-middle A - Left-middle D	0.71318	0.43556
Left-far A - Left-far D	0.00902	0.05810
Centre-close A - Centre-close D	<0.00001 ***	<0.00001 ***
Centre-middle A - Centre-middle D	<0.00001 ***	<0.00001 ***
Centre-far A - Centre-far D	0.00137 *	0.00399 *
Right-close A - Right-close D	0.00001 ***	<0.00001 ***
Right-middle A - Right-middle D	0.00006 ***	0.00074 **
Right-far A - Right-far D	0.00008 ***	0.00021 **

position conditions, we only compared the averted and directed gaze conditions here, as they were the most representative ones for the evaluation of our hypothesis H1. For our two metrics, the normality assumption could not be verified for all our dependent paired samples, thus Student's t-tests or Wilcoxon signed rank tests were run depending on the case. Due to our multiple comparisons, we conducted a Bonferroni correction that changed our target significance level from α=0.05 to α=0.00555. Table A.4 shows the results of these comparisons for the dwell time on the left, and for the fixation count on the right. In the tables, a symbol * indicates a p-value <0.00555, ** a p-value <0.00111, and *** a p-value <0.000111.

These results show an effect of the active agent's position on the dwell time and fixation count results when comparing averted and directed conditions. For seven out nine positions a significant difference was found between these two conditions for this metric, revealing the presence of a stare-in-the-crowd effect; in contrast, for the middle and far left positions, no significant difference was found. Nonetheless, this result is in line with previous studies that discussed the real existence of a stare-in-the-crowd effect across any stimuli positions [Cooper et al., 2013] and any position in the user's field of view [Palanica and Itier, 2011a; Palanica and Itier, 2011b]. In addition, we found that this absence of significant difference between averted and directed condition was due to a larger time spent on the middle/far left field of view on the averted gaze conditions

rather than to a lower one on the directed condition, compared to the results obtained on other positions. This could be explained by a leftward bias of humans during a visual exploration on a scene, as described in the literature [Bourgeois et al., 2018; Foulsham et al., 2013; Ossandón et al., 2014]. Finally, this difference on the left may also have been caused by our experimental stimuli. Indeed, in our experiment the averted gazes of the virtual crowd were always towards a distractor – our virtual speaker – positioned at the left of the user, meaning that the majority of the virtual crowd was looking in that direction. Yet, in their study about the stare-in-the-crowd effect, Palanica et Itier [2011b] found a congruency effect of the averted gazes on the user's gaze behavior, in the sense that active agents whose positions were in the direction signaled by averted gazes were detected faster. Similarly, Sun et al. [2017] also found an effect of the perceived direction of the gaze of the virtual crowd on users' gaze behavior, where users tend to look towards the same direction that they perceive when the majority of the crowd is looking towards one particular direction – in our case to the left.

We also wanted to test if the active agent's position could have affected other metrics than dwell time and fixation count. We found an effect for the first fixation time on the trials where the active agent was in the centre – without distinction of depth *i.e.* 6 trials in total for each gaze condition (3 positions by left/central/right zone * 2 repetitions for each user). For these data samples, the normality assumption could not be verified for all our dependent paired samples, thus Student's t-tests or Wilcoxon signed rank tests were run depending on the case. Due to our multiple comparisons, we conducted a Bonferroni correction that changed our target significance level from α=0.05 to α=0.016. Table A.5 shows the results of these first fixation time comparisons, with one column for each gaze comparison studied, one line for each position zone – left/central/right, and one final line with the p-value previously obtained with the global data without position distinction. In this table, a symbol * indicates a p-value <0.0160, ** a p-value <0.0033, and *** a p-value <0.0003.

These comparison results give new insights on the *first fixation time* metric, and provide for new interpretations about the effect of gaze conditions on it. First the data where all positions are gathered show no significant differences between any gaze condition, as well as the results considering only left or right positions. However, data related to central positions reveal different results with: 1) the presence of significant differences for the comparisons between averted and directed gaze conditions (A vs. D), averted and directed-then-averted ones (A vs. DA), and directed and averted-then-directed ones (D vs.

Table A.5 – First fixation time metric - comparisons by pair of gaze conditions and across position zones

Gaze comp.	A vs. D	A vs. DA	D vs. DA	A vs. AD	D vs. AD
Position	p-value	p-value	p-value	p-value	p-value
Left	0.0745	0.0792	0.8816	0.0293	0.4587
Central	0.0027 **	0.0027 **	0.8340	0.3765	0.012 *
Right	0.9018	0.2103	0.2421	0.7535	0.6181
All	0.5312	0.3290	0.6906	0.8712	0.1670

AD), and 2) the absence of significant differences for the other comparisons. Such results are interesting because they are the ones that were expected according to the stare-in-the-crowd effect: indeed, before the first fixation, the three comparisons present in 1) are equivalent to an averted vs. directed gaze comparison, whereas for the two comparisons of 2) gazes are the same ones in both conditions for these two comparisons (two averted, or two directed). These results **confirm the presence of a stare-in-the-crowd effect in virtual reality**, here regarding the results for the first fixation time metric for active agents in **central positions**.

We may have found this effect only in the central position because of visual differences between virtual reality and photographs. Photographs resolution allows for high-quality display of a crowd in a narrow field of view, about 30° for a user looking at a computer screen. In contrast, in our virtual reality setup the total field of view was larger for the user (the 100° of the FOVE headset), but because of resolution issues and the scale 1:1 for the agents used to provide immersion in virtual reality, more space was required for each agent. Therefore, it could explain why previous results are equivalent to our central part results.

Appendix

B Haptic rendering of collision in a virtual crowd

In this Appendix we present the analysis and the result, related to the metrics that we have not details in the main Chapter 4.

B.1 Analysis

This section details the remaining collected data as well as the variables used to evaluate our hypotheses.

As already mentioned, our data was recorded at 45 Hz and at includes the trajectories of participants, as well as the position and orientation of their limbs in the virtual environment using the Xsens sensors and Unity. We also recorded the body poses over time of each character of the virtual crowd. Then, we were able to replay offline the entire trials in order to compute complex operations such as the volume of each collision.

Trajectories

To study $H1$, we compared participants' trajectories through the virtual crowd. To this end, we decomposed the environment into cells based on a Delaunay triangulation [Chew, 1989], the vertices of which were the crowd characters. A trajectory is then represented as a sequence of traversed cells. An example is displayed in Figure B.1, where the displayed trajectory corresponds to the following sequence of cells: $C_{15}C_{18}C_{13}C_{31}C_5C_{30}C_2C_{34}C_4$.

Represented this way, comparison is possible only when the configuration of crowd characters is identical, which is one reason why we ensured to repeat the same configurations through the 3 studied blocks (cf. Section 4.1). In other words, we first grouped trajectories by crowd configuration, and then compared the set of trajectories performed in the same crowd configuration across different conditions.

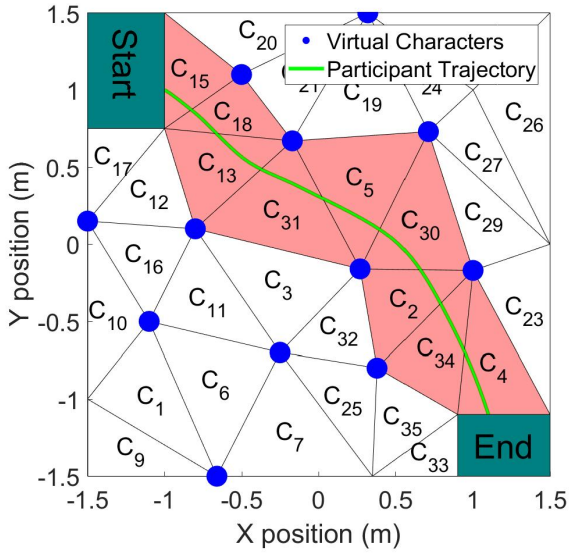


Figure B.1 – Illustration of a participant's trajectory in a crowd, and the decomposition of the environment in cells using Delaunay triangulation [1989].

Comparison was based on the Dice similarity coefficient (DSC) [Sørensen, 1948]. The DSC computes the similarity between two sets A and B according to:

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{B.1}$$

Since our trajectories are sets of traversed cells, two trajectories traversing the same set of cells will be 100% similar. Similarity will decrease with the number of different cells traversed by the participant (occurring in one trajectory and not the other).

Body motions

Navigating in cluttered environments, such as studied in this experiment, requires participants to weave with their body through the crowd. This section presents the data that will be used to analyze body movements when navigating through the virtual crowd to study $H2_1$.

Shoulder rotation. Turning the shoulders is a known strategy for squeezing through narrow openings [Wilmut and Barnett, 2010], i.e., in our case to get between two close characters. To evaluate the effect of haptic rendering on the emergence of such behaviors

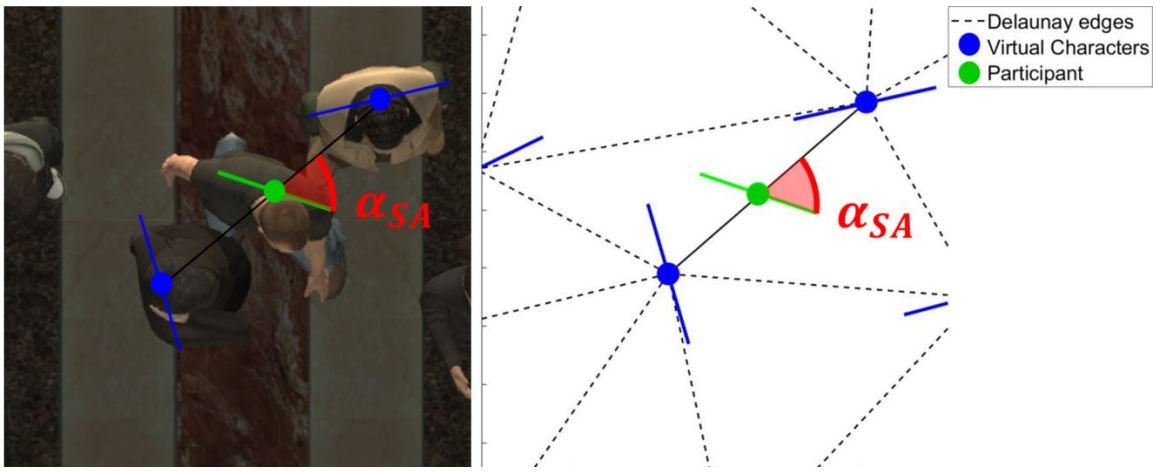


Figure B.2 – Shoulder rotation. Angle $\alpha_{SA} \in [0, 90]°$ is defined between the participants' shoulder-to-shoulder axis and the segment connecting the two virtual characters. Left: top view of the scene. Right: diagram with the Delaunay triangles, the virtual characters, and the participant.

we measured the shoulder orientation at certain critical points of the path. These critical points are the crossing points between the Delaunay cell boundaries (cf. Section B.1) and the participant's trajectories.

More specifically, we computed the angle α_{SA} between the participants' shoulder-to-shoulder axis and the segment connecting the two considered virtual characters, as shown in Figure B.2. This angle provides information about the orientation of the shoulders, and thus the trunk, at the narrowest parts of their path when participants passed between two characters. The larger this angle, the more careful – trying to lower their width at the maximum – participants were when traversing the opening between the two virtual characters.

Walking speed. Beyond the postural analysis introduced in the previous section, we are also interested in the walking speed to analyze whether participants performed the motion task differently according to conditions. To evaluate this parameter only during the navigation, we removed portions of trials where participants were mostly static (e.g., the time during which they were reading the board). To this end, we computed the minimum distance between the participant and the screen, which corresponds to the moment when participants stopped to read the information. We then removed all the frames when the participant's position was less than one step from this position (chosen as 0.74 m for men and 0.67 m for women [Cho et al., 2004]).

Presence and embodiment

Another important aspect of our analysis is its perceptual relevance. In accordance with $H4$, we looked for any difference in the users' feelings of presence and embodiment, comparing the registered subjective perception with and without haptic rendering. Participants answered both questionnaires at the end of each block (Embodiment then Presence), answering each question on a 7-point Likert scale.

Presence. Using an haptic device is generally expected to increase the user's immersion in the virtual world [Krogmeier et al., 2019], as it adds a new sensory feedback, even though it does not always lead to an increase of perceived realism [Slater, 2003]. For this reason, we measured Presence using the Slater-Usoh-Steed (SUS) questionnaire [Usoh et al., 2000] (Table B.5). Each user answered the set of 6 questions, summarized in Table B.5, at the end of each block.

Embodiment. As for Presence, we focused on comparing the sense of embodiment between different blocks to study the influence of the haptic rendering on the perception of the virtual body. We measured embodiment based on the Roth and Latoschik questionnaire [Roth and Latoschik, 2020]. Participants answered Embodiment questionnaires simultaneously with those about Presence.

Statistical analyses

As mentioned, our objective is to understand whether and to what extent users change their behavior in each experimental block. To do so, we analyzed the differences across blocks for all the aforementioned variables. For all dependent variables, we set the level of significance to $\alpha = 0.05$. First, a Shapiro-Wilk test was performed to evaluate whether the distribution of our data followed a normal distribution. If the distribution was not normal, a Friedman test was performed to evaluate the effect of the condition on these variables. Post-hoc comparisons were then performed using a Wilcoxon signed rank test with Bonferroni correction. On the other hand, if the distribution was normal, a one-way analysis of variance (ANOVA) with repeated measures was performed. Greenhouse-Geisser adjustments to the degrees of freedom were applied if the data violated the sphericity assumption. Bonferroni post-hoc tests were used to analyze any significant effects between groups.

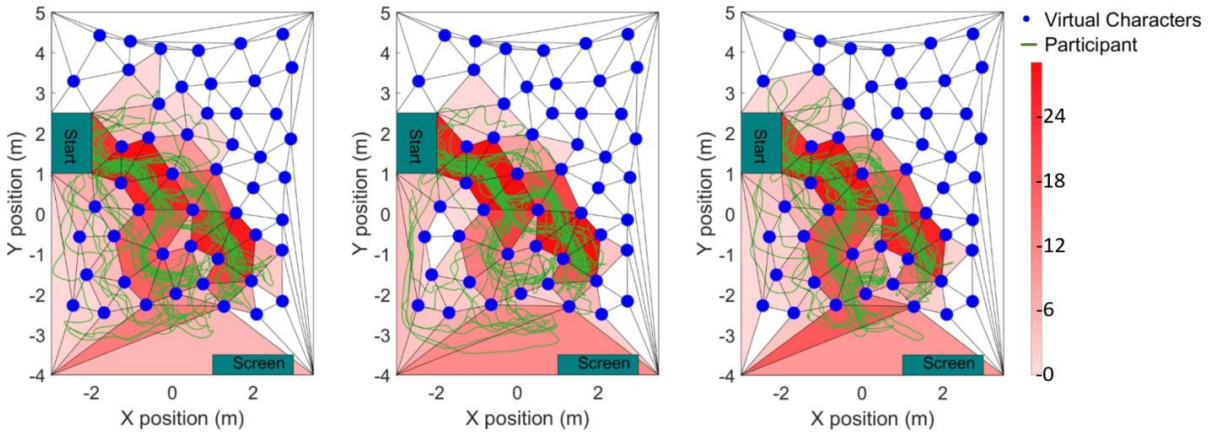


Figure B.3 – Participants' trajectories and Delaunay triangulation for trial T_6 for blocks *NoHaptic1* (left), *Haptic* (middle) and *NoHaptic2* (right). The color-bar represents the number of times participants walked on a triangle.

B.2 Results

This section presents the remaining results of our experiment, starting with the study of $H1$ on the trajectories formed by participants through the virtual crowd. We then explore $H2_1$ and $H2_2$ with respect to the analysis of body movements. Finally, we report the answers to the Presence and Embodiment questionnaires related to $H4$.

Trajectory analysis

Table B.1 shows the results of the Dice similarity measure between all possible pairs of blocks. Similarity ranges from 84.7% (*Nohaptic1* vs. *Haptic* blocks) to 88.5% (*Haptic* vs. *NoHaptic2* blocks). The score is higher for *Haptic* vs. *Nohaptic2* blocks ($88.6 \pm 4.1\%$) and for *Nohaptic1* vs. *Nohaptic2* ($85.9 \pm 4.0\%$).

Because it is difficult to identity from this data only whether the obtained level of similarity is due to natural variety in human behaviors, or to the difference in conditions explored in each block, we propose to measure similarity between paths belonging to the same block as follows. For each block and each configuration, we randomly divided the trajectories into two subsets and computed the Dice similarity score between them. We repeated this process 30 times (which changes the way trajectories are divided into 2 subsets). Performing this process and computing similarity over the 3 blocks resulted into 90 measures of "intra-block similarity". The obtained average value is $81.2 \pm 3.3\%$, that can be compared with the "inter-block similarity" scores presented in Table B.1.

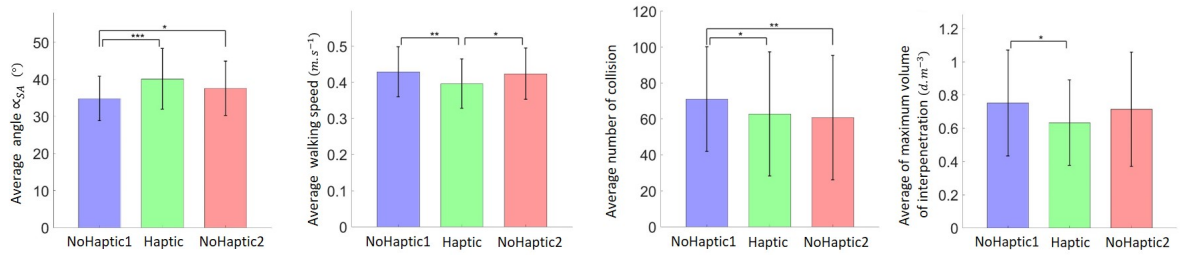
Table B.1 – Similarity measure (Dice) of participant trajectories between all blocks (*NoHaptic1*, *Haptic*, *NoHaptic2*) for all the trials.

Trials	Blocks		
	NoHaptic1 vs. *Haptic*	*Haptic* vs. *NoHaptic2*	*NoHaptic1* vs. *NoHaptic2*
T_1	84.0%	88.6%	85.0%
T_2	88.4%	93.8%	88.3%
T_3	78.1%	93.2%	79.4%
T_4	91.9%	88.7%	90.7%
T_5	88.4%	90.2%	85.3%
T_6	82.8%	85.8%	91.0%
T_7	78.8%	81.6%	82.0%
T_8	85.0%	85.9%	85.3%
$\boldsymbol{T_{all}}$	$\mathbf{84.7 \pm 4.8}$%	$\mathbf{88.6 \pm 4.1}$%	$\mathbf{85.9 \pm 4.0}$%

Our results show that there is no statistical difference between intra-block and *Nohaptic1* vs. *Haptic* blocks similarity measure ($p > 0.05$). There is however a significant difference between intra-block and *Haptic* vs. *Nohaptic2* blocks ($p < 0.01$), as well as intra-block and *Nohaptic1* vs. *Nohaptic2* ($p < 0.05$), where intra-block similarity measures are always lower. Given that similarity measures between pairs of blocks were either as similar or more similar than intra-block similarities, we can conclude that participants chose their path through the crowd similarly, irrespective of the block condition, which supports $H1$. To better illustrate the similarity in navigation paths, Figure B.3 displays all the participants' trajectories and the triangles used to compute the Dice for the specific T_6 configuration.

Body motion

Shoulder rotation. The average amplitude of shoulder rotations α_{SA}, illustrated in Figure B.4.a, was significantly different in each block ($F(2, 44) = 13.0, p < 0.001, \eta_p^2 = 0.37$). In particular, it was significantly higher in the block with haptic rendering ($40.1 \pm 8.2°$), than in the first block without haptic rendering ($34.3 \pm 6.0°$). We remind that a higher α_{SA} angle means that participants made a larger rotation to squeeze between virtual characters, therefore validating the hypotheses $H2_1$. Furthermore, it was also significantly higher in block *NoHaptic2* ($38.7 \pm 3.7°$) than in block *NoHaptic1*, suggesting that participants continued to turn more their shoulders even after haptic rendering was disabled, therefore supporting $H3$.



a) b) c) d)

Figure B.4 – Main significant differences between the three blocks of the experiment (*NoHaptic1*, *Haptic* and *NoHaptic2*): a) amplitude of shoulder rotations (α_{SA}), b) walking speed, c) number of collisions per trial, d) volume of interpenetration. Error bars depict standard deviation of the mean.

Walking speed. We found an effect of haptic rendering ($F(1.56, 34.2) = 7.14, p = 0.005, \eta_p^2 = 0.245$) on participant's average walking speed (Figure B.4.b), where participants' walking speed was on average significantly lower in the*Haptic* block (0.40±0.07 m/s) than in the *NoHaptic1* (0.43±0.07 m/s) and *NoHaptic2* (0.42± 0.07 m/s) blocks. This result therefore supports hypothesis $H2_1$.

Presence and embodiment

The average participant ratings and all the questions for **embodiment** are shown in Tables B.2, B.3 and B.4. We did not find any significant effect of the blocks for *Agency* ($p = 0.438$), *Change* ($p = 0.085$) and *Ownership* ($p = 0.753$). Furthermore, Table B.5 shows the questions and the average participant ratings for **presence**, for which we also did not find a significant effect of the blocks ($p = 0.222$). These results therefore do not support hypothesis $H4$, suggesting that haptic rendering does not improve the sense of presence or the sense of embodiment of participants in virtual reality.

B.3 Discussion

In this Section, we discuss in details the results related to trajectories and Presence and Embodiment.

Table B.2 – *Agency* questionnaire: average participant ratings for the three blocks.

Questions	**blocks**		
	NoHaptic1	*Haptic*	*NoHaptic2*
The movements of the virtual body felt like they were my movements. I felt like I was controlling the movements of the virtual body I felt like I was causing the movements of the virtual body. The movements of the virtual body were in sync with my own movements.	6.1 ± 0.9	6.0 ± 0.8	5.9 ± 0.7

Table B.3 – *Change* questionnaire: average participant ratings for the three blocks.

Questions	**Blocks**		
	NoHaptic1	*Haptic*	*NoHaptic2*
I felt like the form or appearance of my own body had changed. It felt like the weight of my own body had changed. I felt like the size (height) of my own body had changed. I felt like the width of my own body had changed.	3.6 ± 1.3	3.8 ± 1.5	3.3 ± 1.5

Trajectories

In Section B.2, the analysis of the Dice similarity measure showed that haptic rendering did not change the way participants selected their path through the crowd, as stated in hypothesis $H1$. We even found that paths across blocks were "more similar" than within the same block. One possible explanation is given by the way we compose the sets we compare the similarity of, where we assume that paths are independent from participants. Indeed, the intra-block similarity measure required us to split a set of trajectories belonging to the same block and crowd configuration, which resulted into comparing paths performed by different participants. In contrast, the inter-block analysis considered sets that were split according to haptic rendering conditions, thus comparing paths performed by the same group of 23 participants.

In spite of this limitation in our analysis, we consider that paths are similar across

Table B.4 – *Ownership* questionnaire: average participant ratings for the three blocks.

Questions	**Blocks**		
	NoHaptic1	*Haptic*	*NoHaptic2*
It felt like the virtual body was my body. It felt like the virtual body parts were my body parts. The virtual body felt like a human body. It felt like the virtual body belonged to me.	4.9 ± 1.4	5.1 ± 1.2	5.0 ± 1.2

Table B.5 – Slater-Usoh-Steed (SUS) questionnaire [2000] and average participant ratings for the three blocks.

Questions	**Blocks**		
	NoHaptic1	*Haptic*	*NoHaptic2*
I had a sense of being there in the train station. There were times during the experience when... the train station was the reality for me... The train station seems to me to be more like... I had a stronger sense of... I think of the train station as a place in a way similar to other places that I've been today. During the experience I often thought that I was really standing in the train station.	5.2 ± 0.9	5.2 ± 1.2	5.0 ± 1.1

blocks. One can describe human motion as a trajectory resulting from a perception-action loop [Gibson, 1958; Warren, 1998]. Depending on the tasks, the loop is a multi-modal one, meaning that different senses are used to control motion. However, in the context of walking, several studies [Patla, 1997; Warren, 1998] have shown that vision is the most used perceptual input to navigate to the goal. Such statements hold in our case, where a major difference with previous work is the higher density of obstacles. Nevertheless, assuming that tactile feedback may affect path selection, it would have been probable that some participants reversed their course after a collision has been rendered, which was not observed.

Embodiment and presence

In contrast with our hypothesis $H4$, we did not find any significant change in terms of user's perceived senses of embodiment and presence when experiencing haptic feedback. This result is quite surprising, as we did find significant effects in other measurements,

suggesting that participants took different actions when provided with haptic sensations of contact. An explanation for this result could lie in the fact that users already registered high embodiment and presence levels without experiencing haptic feedback in the first condition (*NoHaptic1*), leaving little room for improvement in the *Haptic* condition. Another possibility is that vibrotactile feedback is not suited to render collisions in crowds, although there are several examples of this type of feedback being used to render similar events [Bimbo et al., 2017; Devigne et al., 2020]. Finally, a last explanation could be the location and number of our haptic devices. Employing a higher number of bracelets spread throughout the body might better render the target contact sensations. All these considerations will drive our future work, as discussed in Chapter 4, Section 4.5.

List of Figures

List of Tables

Université de Rennes

Titre : titre (en français).....

Mot clés : de 3 à 6 mots clefs

Résumé : Eius populus ab incunabulis primis ad usque pueritiae tempus extremum, quod annis circumcluditur fere trecentis, circumm urana pertulit bella, deinde aetatem ingressus adultam post multiplices bellorum aerumnas Alpes transcendit et fretum, in iuvenem erectus et virum ex omni plaga quam orbis ambit inmensus, reportavit laureas et triumphos, iamque vergens in senium et nomine solo aliquotiens vincens ad tranquilliora vitae discessit. Hoc inmaturo interitu ipse quoque sui pertaesus excessit e vita aetatis nono anno atque vicensimo cum quadriennio imperasset. natus apud Tuscos in Massa Veternensi, patre Constantio Constantini fratre imperatoris, matreque Galla. Thalassius vero ea tempestate praefectus praetorio praesens ipse quoque adrogantis ingenii, considerans incitationem eius ad multorum augeri discrimina, non maturitate vel consiliis mitigabat, ut aliquotiens celsae potestates iras principum molliverunt, sed adversando iurgandoque cum parum congrueret, eum ad rabiem potius evibrabat, Augustum actus eius exaggerando creberrime docens, idque, incertum qua mente, ne lateret adfectans. quibus mox Caesar acrius efferatus, velut contumaciae quoddam vexillum altius erigens, sine respectu salutis alienae vel suae ad vertenda opposita instar rapidi fluminis irrevocabili impetu ferebatur. Hae duae provinciae bello quondam piratico catervis mixtae praedonum.

Title: titre (en anglais).....

Keywords: de 3 à 6 mots clefs

Abstract: Eius populus ab incunabulis primis ad usque pueritiae tempus extremum, quod annis circumcluditur fere trecentis, circumm urana pertulit bella, deinde aetatem ingressus adultam post multiplices bellorum aerumnas Alpes transcendit et fretum, in iuvenem erectus et virum ex omni plaga quam orbis ambit inmensus, reportavit laureas et triumphos, iamque vergens in senium et nomine solo aliquotiens vincens ad tranquilliora vitae discessit. Hoc inmaturo interitu ipse quoque sui pertaesus excessit e vita aetatis nono anno atque vicensimo cum quadriennio imperasset. natus apud Tuscos in Massa Veternensi, patre Constantio Constantini fratre imperatoris, matreque Galla. Thalassius vero ea tempestate praefectus praetorio praesens ipse quoque adrogantis ingenii, considerans incitationem eius ad multorum augeri discrimina, non maturitate vel consiliis mitigabat, ut aliquotiens celsae potestates iras principum molliverunt, sed adversando iurgandoque cum parum congrueret, eum ad rabiem potius evibrabat, Augustum actus eius exaggerando creberrime docens, idque, incertum qua mente, ne lateret adfectans. quibus mox Caesar acrius efferatus, velut contumaciae quoddam vexillum altius erigens, sine respectu salutis alienae vel suae ad vertenda opposita instar rapidi fluminis irrevocabili impetu ferebatur. Hae duae provinciae bello quondam piratico catervis mixtae praedonum.