Télé-immersion 3D basée sur des caméras 360° et des casques de réalité étendue

Clément Dluzniewski

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : I-MEP² - Ingénierie - Matériaux, Mécanique, Environnement, Energétique, Procédés, Production
Spécialité : GI - Génie Industriel : conception et production
Unité de recherche : Laboratoire des Sciences pour la Conception, l'Optimisation et la Production de Grenoble

Télé-immersion 3D basée sur des caméras 360° et des casques de réalité étendue

3D Tele-Immersion based on 360° Cameras and Extended Reality Headsets

Présentée par :

Clément DLUZNIEWSKI

Direction de thèse :

Frédéric NOEL PROFESSEUR DES UNIVERSITES, GRENOBLE INP - UGA	Directeur de thèse
Jérémie LE GARREC CEA	Co-encadrant de thèse

Rapporteurs :

Géraldine MORIN
PROFESSEURE DES UNIVERSITES, Toulouse INP
Frédéric MERIENNE
PROFESSEUR DES UNIVERSITES, ENSAM CER Cluny

Thèse soutenue publiquement le **23 septembre 2024**, devant le jury composé de :

Georges DUMONT, PROFESSEUR DES UNIVERSITES, Ecole Normale Supérieure de Rennes	Président
Frédéric NOEL, PROFESSEUR DES UNIVERSITES, Grenoble INP - UGA	Directeur de thèse
Géraldine MORIN, PROFESSEURE DES UNIVERSITES, Toulouse INP	Rapporteure
Frédéric MERIENNE, PROFESSEUR DES UNIVERSITES, ENSAM CER Cluny	Rapporteur
Laurence NIGAY, PROFESSEURE DES UNIVERSITES, Université Grenoble Alpes	Examinatrice

Invités :

Jérémie LE GARREC
CHARGE DE RECHERCHE, CEA
Claude ANDRIOT
CHARGE DE RECHERCHE, CEA

TÉLÉ-IMMERSION 3D BASÉE SUR DES CAMÉRAS 360° ET DES CASQUES DE RÉALITÉ ÉTENDUE

CEA-List - Université Grenoble Alpes
France

Clément Dluzniewski

2024

Table des matières

Liste des tableaux

Table des figures

Résumé

Dans les environnements professionnels contemporains, le travail est souvent dispersé sur différents lieux géographiquement éloignés. La réunion des collaborateurs en présentiel pouvant s'avérer complexe, les professionnels se reposent aujourd'hui sur les technologies de l'information et de la communication pour organiser les interactions. La télé-immersion s'inscrit dans cette continuité de technologies avec l'ambition de rapprocher les individus séparés géographiquement comme s'ils étaient présents dans un même lieu. Cette thèse propose de réaliser un système de télé-immersion original basé sur une caméra 360° et des casques de réalité étendue. Ce système est conçu pour que des utilisateurs dans un lieu d'intérêt puissent ramener auprès d'eux des utilisateurs distants simplement en posant une caméra 360°. Grâce à la réalité étendue, les utilisateurs distants sont comme téléportés sur le lieu d'intérêt. Le système est spécifiquement développé pour répondre au besoin de nouvelles technologies d'enseignement à distance, afin que des enseignants puissent dispenser des cours immersifs à des étudiants chez eux.

Le premier verrou pour atteindre un tel système consiste à gérer le point de vue de multiples utilisateurs avec une seule caméra 360°, tout en augmentant le sentiment de coprésence. En effet, chaque utilisateur distant ayant le point de vue de la caméra, tous se retrouvent localisés au même endroit sur le lieu d'intérêt. Le second verrou est de développer des interactions avec les données de la caméra 360°. Nous souhaitons particulièrement proposer aux utilisateurs distants de naviguer librement sur le lieu d'intérêt. Le problème est alors de trouver une représentation du lieu capable de générer plusieurs points de vue et qui peut être capturée avec une caméra 360° statique.

Le manuscrit présente quatre contributions : un nouveau cadre théorique de la télé-immersion et trois versions de systèmes de télé-immersion basés sur une unique caméra 360° statique. La première exploite uniquement des images 360° sans informations 3D, la seconde intègre des informations 3D aux images 360° sous forme de cartes de profondeur, et la dernière profite d'une nouvelle représentation 3D 360°. Cette dernière version est évaluée avec une expérience utilisateur visant à montrer que le sentiment de présence qu'elle suscite est plus grand qu'avec la simple diffusion de la vidéo 360° capturée par la caméra.

Publications

La réalisation de la thèse a conduit à la diffusion des publications suivantes :

— Clément Dluzniewski, Jérémie Le Garrec, Claude Andriot, et Frédéric Noël : Light VR Client for Point Cloud Navigation with 360° Images. *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022.

— Clément Dluzniewski, Jérémie Le Garrec, Claude Andriot, et Frédéric Noël : Interacting with 3D Avatars and Laser Pointers in a 360° Image. *EuroXR*, 2022.

— Clément Dluzniewski, Hakim Chekirou, Jérémie Le Garrec, Claude Andriot, et Frédéric Noël : 3D Reconstruction for Tele-Immersion in 360° Live Stream. *ICAT-EGVE*, 2023.

Chapitre 1

Introduction

Faciliter la collaboration à distance entre les humains est un défi récurrent dans l'histoire. Au fil des nouvelles technologies, les manières de travailler avec des personnes distantes ont évoluées, allant du téléphone à la visioconférence en passant par les e-mails. Cependant, un point fait généralement consensus : ces techniques ne sont pas comparables à la collaboration en face à face traditionnelle. Cela est particulièrement vrai pour les enseignants et les étudiants qui préfèrent que les cours se déroulent dans une salle où tous sont physiquement présents. Mais certaines circonstances rendent la réunion des participants en présentiel impossible. Quand ces contraintes ne peuvent pas être levées, la télé-immersion prend tout son sens. Avec un système de télé-immersion idéal, les différents collaborateurs se retrouvent sur un lieu commun sans s'y rendre physiquement et retrouvent les avantages d'une collaboration en face à face. Dans ce chapitre, nous donnons un aperçu de la télé-immersion et décrivons ses avantages pour l'enseignement à distance. Puisque nous souhaitons développer notre propre système de télé-immersion, ce chapitre introduit le système vers lequel nous désirons tendre. Enfin, les questions de recherche avec les hypothèses et la structure du manuscrit sont présentées.

1.1 De la Fiction à la Réalité

De tout temps, œuvres de fiction et progrès scientifique et technologique se sont mutuellement influencés. L'appel à la science semble lié à un attrait pour la science-fiction (Fleischmann et Templeton, 2009) et il n'est pas étonnant que de réelles technologies trouvent leurs origines dans des technologies fictives (Mair, 2013; Raitt *et al.*, 2001). Une technologie faisant partie du paysage des univers de science-fiction est la téléportation. Dans ces univers, la téléportation est le transport d'un objet d'un endroit à un autre sans passer par l'espace physique qui les sépare. Avoir un système de téléportation est idéal pour des personnes géographiquement éloignées qui souhaitent collaborer en

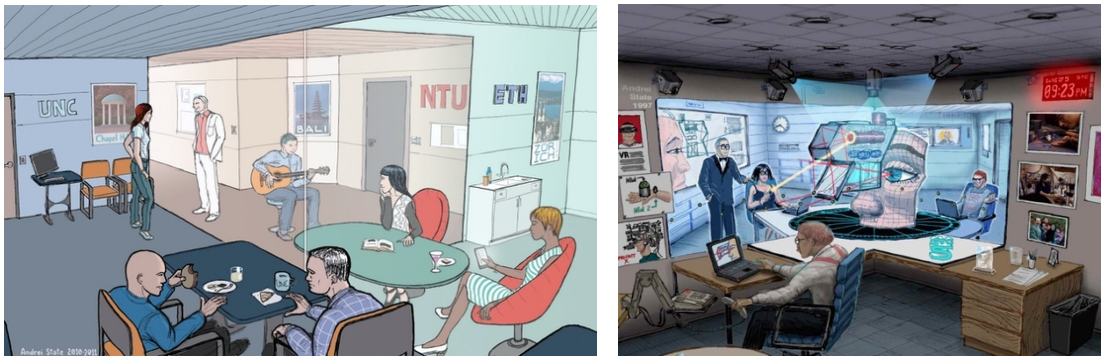


FIGURE 1.1 – Exemples de systèmes de télé-immersion fictifs. Gauche : Télé-immersion multi-salles du *BeingThere Centre* (Fuchs *et al.*, 2014). Droite : *Office of the Future* (Raskar *et al.*, 1998).

face à face car elle offre un trajet vers le lieu de collaboration bien plus rapide et bien moins coûteux. Une telle technologie paraît physiquement impossible en réalité, son impact est donc certainement mineur sur les technologies de collaboration à distance. Mais une autre famille de technologies plus crédible a pu engendrer de réels systèmes de collaboration à distance. Ces technologies, qualifiées de télé-immersion, entendent réunir des personnes éloignées sur un site commun sans les déplacer physiquement. De manière intéressante, les propositions de télé-immersion dans les œuvres littéraires et audiovisuelles empruntent largement à la réalité étendue (Fast-Berglund *et al.*, 2018). Par exemple, les personnages de *Star Trek* bénéficient du *Holodeck*, une salle semblable à un CAVE créant des environnements virtuels dans lesquels ils sont immergés. Dans le roman *Neuromancien*, les personnages s'équipent d'un casque *deck* pour s'isoler du monde réel et rejoindre un environnement virtuel. Ces systèmes, basés sur la réalité virtuelle, représentent les systèmes de télé-immersion où l'utilisateur ne perçoit plus son environnement réel immédiat pour être transporté dans un autre environnement. Les différentes personnes désirant collaborer se réunissent dans cet environnement virtuel. D'autres systèmes de télé-immersion vont plutôt simuler la présence d'une personne éloignée en réalité augmentée pour donner l'impression qu'elle est physiquement sur le lieu. C'est le cas dans *Star Wars* où les participants à distance d'une réunion sont présentés sous forme d'hologramme, ou dans *Kingsman* avec des lunettes de réalité augmentée. Les systèmes de collaboration à distance contemporains (Schäfer *et al.*, 2023) ont probablement été inspirés par des systèmes de télé-immersion fictifs et exercé une influence sur cet imaginaire.

La littérature scientifique est aussi à la source d'idées de futures technologies de télé-immersion. Les chercheurs du *BeingThere Centre*, un effort de recherche international conjoint entre l'Université technologique de Nanyang à Singapour, l'ETH Zürich en Suisse et l'Université de Caroline du Nord aux États-Unis, ont dirigé de nombreux

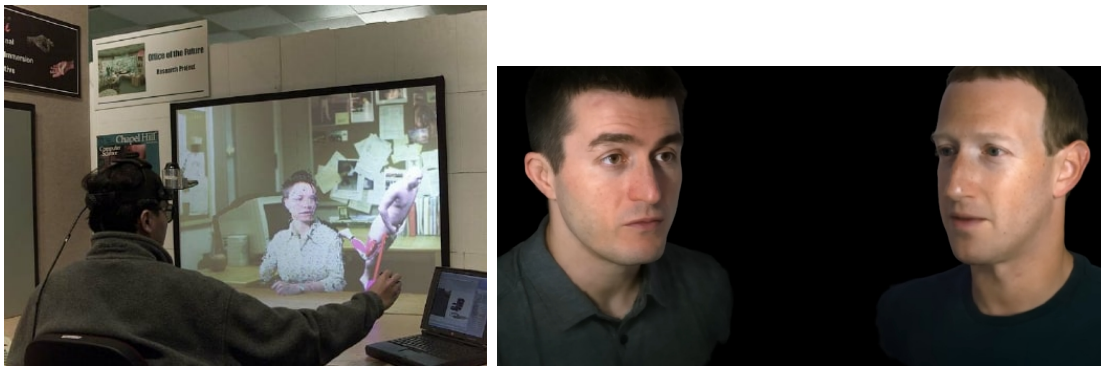


FIGURE 1.2 – Exemples de systèmes de télé-immersion réels. Gauche : (Towles *et al.*, 2003). Droite : Conversation avec des *Pixel Codec Avatars* (Ma *et al.*, 2021) (extrait vidéo [1]).

projets de télé-immersion (Fuchs *et al.*, 2014). Parmi ces projets, ils ont notamment imaginé un concept de télé-immersion multi-salles où des sites distants sont reliés virtuellement comme s'ils partageaient le même espace grâce à des écrans transparents (figure 1.1 gauche). Le projet *Office of the Future* (Raskar *et al.*, 1998) dans les années 1990 a aussi inspiré de nombreux systèmes. L'idée était d'avoir un espace de travail avec des collaborateurs à distance sur un même bureau (figure 1.1 droite). Dans cet esprit, Cisco lance en 2006 le *TelePresence System 3000*, un système commercial de salle de réunion où les participants distants sont filmés et affichés avec de larges écrans autour de la table (Szigeti *et al.*, 2009). Les grands projets à l'échelle nationale ont été des jalons dans l'histoire de la télé-immersion. Le plus important d'entre-eux est la *National Tele-Immersion Initiative* (*NTII*) développée entre les années 1990 et 2000 à travers les États-Unis pour concevoir une infrastructure de collaboration entre des personnes à différents endroits du pays (Sadagic *et al.*, 2001). Financé par la *National Science Foundation*, l'objectif était de faciliter la collaboration dans la recherche scientifique ou la formation à distance, mais aussi la télémédecine ou la chirurgie téléassistée. Ce projet a donné naissance à des systèmes tels que celui de (Towles *et al.*, 2003) qui simule la présence d'utilisateurs distants en taille réelle en face d'un bureau grâce à des caméras stéréoscopiques et des écrans 3D (figure 1.2 gauche). Le *CAVE Research Network* a aussi contribué au développement de plusieurs systèmes de télé-immersion centrés sur l'immersion dans un CAVE (Leigh *et al.*, 1999). D'autres projets d'envergure sont aussi lancés en Europe, comme en Allemagne avec le centre de recherche collaborative qui réalise en 2010 des projets de chirurgie peu invasive, de téléfabrication et de télémaintenance (Bauernschmitt *et al.*, 2010), en Grèce en 2016 avec la proposition d'un système de télé-immersion complet (Zioulis *et al.*, 2016), ou en Slovaquie en 2015 avec le projet d'Infrastructure Nationale de Téléprésence (*NTI*) pour le transfert de technologie

1. `https://youtube.com/watch?v=MVYrJJNdrEg`

(Jakab *et al.*, 2016). Aujourd'hui, la technologie qui semble le plus à même d'être un tournant décisif pour le futur de la télé-immersion est le *Pixel Codec Avatar* (Ma *et al.*, 2021). Cette technologie permet, avec un scan du visage capturé au préalable, d'offrir un avatar photoréaliste de la personne, animé par les mouvements de son visage (figure 1.2 droite). La généralisation de cette solution serait une opportunité pour rassembler des utilisateurs avec des avatars aussi vrais que nature dans des environnements virtuels. Cette thèse propose d'enrichir la liste des systèmes actuels en réalisant un système de télé-immersion original facilement déployable, basé sur une caméra 360° et des casques de réalité étendue.

1.2 Contexte

L'enseignement à distance, aujourd'hui appelé aussi enseignement en ligne ou e-éducation, est un mode d'enseignement pour les étudiants qui ne sont pas dans un environnement éducatif physique comme une salle de classe. Des cours par correspondance aux cours en ligne ouvert à tous (*MOOC*), l'enseignement à distance permet à des personnes de s'instruire sans se rendre sur les lieux d'apprentissage traditionnels. Mais jusqu'aux technologies récentes, il était difficile pour un étudiant d'assister à un cours classique, où l'étudiant et l'enseignant interagissent de manière synchrone, sans être physiquement sur le lieu. Aujourd'hui, l'enseignement à distance peut s'appuyer sur les systèmes de télé-immersion pour qu'étudiants et enseignants puissent se rejoindre comme s'ils étaient dans le même lieu.

Historiquement, trois grandes technologies ont dominé l'enseignement à distance : l'impression, l'audiovisuel et le numérique (Bozkurt, 2019). L'utilisation de plus en plus fréquente du numérique a permis un plus large accès à l'enseignement à distance sous différentes formes comme les *MOOC* ou les retransmissions vidéos des cours d'universités (Anderson et Rivera Vargas, 2020). Les périodes de confinement vécues ces dernières années ont renforcé la mise en place de ces pratiques d'enseignement en ligne. En particulier, pour maintenir la formation des étudiants, cet épisode a généralisé l'adoption de l'enseignement à distance sous forme de visioconférence. Dans sa configuration standard, étudiants et professeurs se retrouvent sur une plateforme de visioconférence en activant les flux vidéos de leurs webcams. Cependant, la plupart des étudiants semblent moins satisfaits et ont de moins bons résultats académiques avec cette manière d'étudier par rapport à l'apprentissage traditionnel en face à face (Roth *et al.*, 2020). Ceux-ci ressentent bien moins d'engagement dans le cours et pas d'amélioration dans leurs participations (Serhan, 2020). On peut supposer que le simple visionnage vidéo de l'enseignant sur un écran n'est pas suffisant pour qu'un étudiant s'implique et se concentre sur le cours. Les professeurs aussi peuvent avoir des difficultés à poursuivre le cours en raison de la difficulté d'évaluer l'état de concentration des étudiants. Les interactions

informelles entre les étudiants, qui peuvent renforcer la compréhension du cours, ne sont pas possibles non plus avec ces plateformes. Enfin, la tenue de travaux pratiques, où il y a besoin de manipuler des éléments, n'est pas concevable avec une interface de visioconférence limitée. Néanmoins, dans de nombreux cas, il peut être intéressant de développer des outils plus élaborés pour tenir des cours à distance, comme l'impossibilité pour l'étudiant de se déplacer ou l'incapacité d'accueil de l'établissement.

Une piste pour régler les problèmes des cours à distance en visioconférence serait d'utiliser la réalité virtuelle (Speidel *et al.*, 2023). Ces technologies immersives permettraient idéalement d'immerger professeurs et étudiants dans une salle de classe virtuelle et d'avoir un cours comme s'il se déroulait en présentiel. Combinés avec des jumeaux numériques, les professeurs peuvent aussi développer des sessions de travaux pratiques où un étudiant pourrait manipuler une copie virtuelle d'une machine comme s'il la manipulait en réel. C'est dans ce contexte qu'a été développé le projet JENII (Jumeaux d'Enseignement Numériques, Immersifs et Interactifs) qui vise à créer des formations à distance immersives et collaboratives, dans lequel le CEA est impliqué avec les Arts et Métiers, le CESI et le CNAM. Une direction investiguée est alors d'utiliser une caméra 360°, au lieu d'une simple webcam, afin de filmer le professeur donnant cours sur site et d'immerger en réalité virtuelle les étudiants à distance. Des plateformes d'enseignement à distance basées sur la diffusion en direct du flux 360° dans des dispositifs de réalité virtuelle existent déjà pour immerger un étudiant dans une salle de classe ou une salle de travaux pratique. Ceux-ci ont montré un plus grand sentiment de présence spatiale et sociale par rapport à la visioconférence classique (Gandsas *et al.*, 2023; Orduna *et al.*, 2022). Le sujet de thèse s'inscrit dans cette continuité sur le thème de la télé-immersion 3D dont la finalité est la création d'une plateforme de télé-immersion permettant à un enseignant de réunir des élèves à distance dans la salle de cours.

1.3 Système de Télé-Immersion *Idéal*

Le cœur des chapitres suivants sera consacré à la recherche et au développement d'un système de télé-immersion idéal. Le système idéal auquel nous souhaitons parvenir possède un ensemble de propriétés qu'aucun système actuellement développé ne présente simultanément.

1.3.1 Propriétés

Premièrement, notre système pour une télé-immersion idéale permettra de ramener des utilisateurs distants sur un site d'intérêt réel auprès d'utilisateurs physiquement présents. Les utilisateurs sur le lieu verront des représentations des utilisateurs distants

pour simuler leurs présences. Les utilisateurs distants seront eux virtuellement transportés sur le lieu d'intérêt. Pour un usage général, le lieu d'intérêt peut être une salle de réunion pour collaborer, un site industriel pour vérifier les normes d'une installation ou une salle de travaux pratiques pour réaliser des manipulations.

Notre système devra se baser sur un dispositif permettant de capturer l'ensemble du lieu, et pas seulement les personnes sur site ou un objet d'intérêt particulier. Dans l'état de l'art, les propositions se basent principalement sur des dispositifs qui capturent uniquement les représentations des personnes plutôt que le lieu dans son intégralité. Ces systèmes permettent alors de rassembler les utilisateurs sur un environnement virtuel (Viola *et al.*, 2023; Zioulis *et al.*, 2016) ou de visualiser les utilisateurs distants localement sur son propre site (Córdova-Esparza *et al.*, 2019; Orts-Escolano *et al.*, 2016). Mais, ces dispositifs ne permettent pas d'immerger un utilisateur comme s'il était sur un site distant réel. Nous avons alors choisi d'utiliser un dispositif facilement transportable pour acquérir l'environnement. L'intérêt d'un tel dispositif est qu'il permet d'avoir un système de télé-immersion nomade avec lequel on peut immerger des utilisateurs distants dans un nouvel endroit simplement en transportant le dispositif dans ce nouveau lieu (pas besoin de concevoir un dispositif spécifiquement pour le lieu). Nous nous sommes alors orientés vers le plus simple des appareils pouvant capturer une vue globale du lieu : la caméra 360°. Aujourd'hui accessible et bien étudiée, la caméra 360° assure la mobilité de notre système et sa généralisation au plus grand nombre. Mais nous ambitionnons un nouvel usage de la caméra 360° : capturer le lieu de sorte qu'un utilisateur distant puisse se déplacer à l'intérieur librement. Un système où l'utilisateur distant se déplace en fonction de la position d'un utilisateur sur site est techniquement possible à l'aide d'une caméra 360° embarquée. Cependant, nous tenons à offrir à l'utilisateur distant un point de vue indépendant qui n'est pas asservi à celui d'un utilisateur sur site. Cette décision est motivée par le fait que la vue indépendante semble être préférée par un utilisateur distant par rapport à une vue dépendante de l'utilisateur sur site. Cette vue indépendante aide notamment à réduire le temps d'exécution d'une tâche (Kim *et al.*, 2018) et à avoir une meilleure compréhension des actions de l'utilisateur sur site sans diminution du niveau de conscience (Tait et Billinghurst, 2015). Cette navigation libre permet également de supporter simultanément la représentation de plusieurs utilisateurs distants. En effet, à travers une vidéo 360°, on ne peut voir le site que du point de vue de la caméra, il est donc impossible de s'y déplacer si la caméra n'est pas physiquement déplacée. Filmer avec une simple caméra 360° implique donc que les utilisateurs distants sont tous virtuellement à la même position sur le site. Mais s'ils sont tous à la même position, il est inutile de les représenter individuellement avec des avatars car leurs représentations seraient toutes entassées et indistinguables. La communication, en particulier la communication non-verbale, est donc plus difficile entre

un utilisateur sur site et un utilisateur distant, ou même entre les utilisateurs distants. Une autre possibilité pour notre système est de monter la caméra 360° sur une base mobile pouvant être contrôlée par un utilisateur distant (Kim *et al.*, 2023; Jones *et al.*, 2020a; Heshmat *et al.*, 2018; Ogi et Fueki, 2017) choisissant ainsi librement son point de vue en déplaçant physiquement la caméra. Cependant, ce choix de système aurait des implications pour les utilisateurs distants. Soit le système impose aux utilisateurs distants de tous partager la même caméra 360° sur le site d'intérêt. Dans ce cas, les utilisateurs distants continuent de partager le même point de vue (pas de représentations pour chaque utilisateur individuel). Ceux-ci perdent aussi en liberté dans leurs choix du point de vue car un compromis doit être trouvé parmi eux pour décider où déplacer la plateforme mobile. Soit le système propose une caméra 360° individuelle pour chaque utilisateur distant. Mais ce cas irait en contradiction avec notre volonté de se reposer sur un dispositif de capture simple pour créer un système mobile. La télé-immersion dans un nouveau lieu nécessiterait de transporter autant de caméras 360° mobiles que d'utilisateurs distants. De plus, une utilisation de plusieurs caméras 360° pourrait être plus judicieuse en les disséminant à travers le lieu de manière statique pour obtenir une reconstruction 3D complète (da Silveira *et al.*, 2022). La reconstruction 3D permettant la même liberté de navigation que le contrôle d'une caméra 360° pour chaque utilisateur distant. Nous avons alors opté pour capturer le site d'intérêt avec une unique caméra 360°.

Notre système de télé-immersion doit aussi permettre de télé-immerger des utilisateurs sur des lieux dynamiques. Ce type de lieu implique une capacité de créer et de transmettre une représentation en temps réel afin que l'utilisateur distant le perçoive dans un délai suffisamment court pour assurer l'interactivité (Irlitti *et al.*, 2023; Kolkmeier *et al.*, 2018; Adcock *et al.*, 2013; Maimone et Fuchs, 2012). L'interaction avec le lieu incluant la manipulation physique d'objets ou la communication avec une personne sur site. Ceci nous a donc incité à nous servir de la caméra 360° de manière statique : la caméra 360° est posée sur le lieu et reste immobile. En effet, pour supporter une navigation libre des utilisateurs distants, une solution est de reconstruire le site en 3D. Si l'environnement est statique, l'utilisateur peut être immergé dans sa représentation 3D en déplaçant simplement le dispositif d'acquisition à travers le site pour le reconstituer en 3D dans son ensemble (Stotko *et al.*, 2019a; Niessner *et al.*, 2013). Mais si l'environnement est dynamique, des éléments peuvent être modifiés, apparaître ou disparaître, hors du champ de vision de la caméra. La reconstruction 3D ne pourra alors pas être une copie réelle du site. Avec une caméra 360° fixe, on peut espérer que l'ensemble de la zone d'intérêt du site soit visible, et donc avoir une reconstruction 3D à jour.

Enfin, les utilisateurs distants doivent se sentir présents sur le lieu où ils sont télé-immergés, tandis que les utilisateurs sur site doivent avoir l'impression que les utilisateurs distants sont avec eux. Nous avons alors utilisé la réalité étendue pour les dispositifs d'immersion. En particulier, la réalité virtuelle pour immerger les utilisateurs distants, et la réalité mixte pour que les utilisateurs sur site voient les utilisateurs distants en surimpression. Les utilisateurs distants sont immergés dans les données de la caméra 360° (dans lesquels ils peuvent se déplacer librement), et sont représentés par des avatars 3D (génériques ou reconstruits à l'avance) contrôlés avec le dispositif de réalité virtuelle (position des manettes qui animent les mains...). Les utilisateurs sur site, capturés par la caméra 360°, seront représentés par des avatars 3D créés à la volée à partir des données de la caméra.

1.3.2 Proposition Générale

Les contraintes posées permettent de dessiner les contours du système de télé-immersion vers lequel nous souhaitons tendre. Le scénario est le suivant : des utilisateurs sur un site d'intérêt veulent amener auprès d'eux des utilisateurs distants. Les utilisateurs sur site ont l'impression que ceux à distance sont physiquement avec eux sur le site. Les utilisateurs distants sont comme transportés sur le site d'intérêt. Pour cela, les utilisateurs sur site placent au centre du lieu une caméra 360° immobile. Celle-ci capture l'ensemble du site en 3D et transmet ces informations aux utilisateurs distants qui vont pouvoir naviguer librement. Les utilisateurs distants se sentent présents sur le site d'intérêt grâce à des casques de réalité virtuelle qui les immergent totalement dans une représentation photoréaliste 3D. Du côté des utilisateurs sur site, on voit les utilisateurs distants en surimpression dans l'environnement grâce à des casques de réalité mixte. Si un utilisateur distant est à une certaine position dans la représentation 3D du site, alors les utilisateurs locaux verront son avatar à cette même position sur le site réel. La figure 1.3 illustre ce scénario.

Le système proposé est inspiré par des systèmes de l'état de l'art. La télé-immersion 360° traditionnelle consiste simplement à diffuser aux utilisateurs distants la vue de la caméra 360° en réalité virtuelle. Nous proposons un système plus sophistiqué combinant caméra 360° et réalité mixte, largement inspiré par AVT (Rhee *et al.*, 2020). AVT est un système dans lequel un utilisateur local accueille un utilisateur distant dans son environnement qui est filmé par une seule caméra 360° statique. L'utilisateur distant reçoit le flux vidéo 360° filmé par la caméra en temps réel, tandis que l'utilisateur sur site voit un avatar de l'utilisateur distant en surimpression grâce à un casque de réalité augmentée. Une idée similaire est présente dans (Pece *et al.*, 2013) sans réalité mixte, mais pour des visioconférences. Cependant, aucune information 3D n'est capturée dans AVT, l'utilisateur distant est donc fixé à la position de la caméra, l'empêchant ainsi

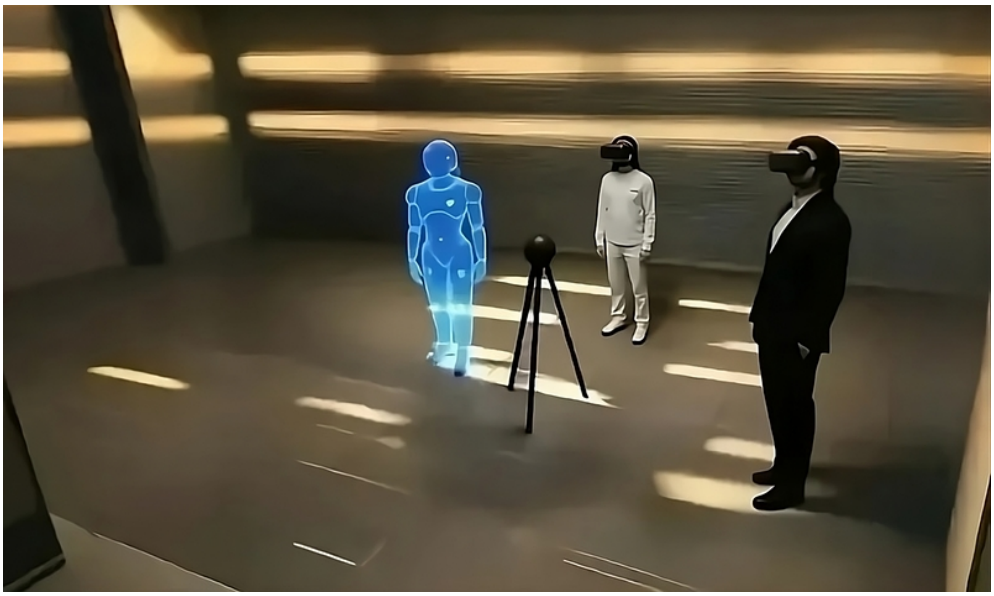


FIGURE 1.3 – Illustration conceptuelle de notre système de télé-immersion (image créée par intelligence artificielle générative). Une caméra 360° est placée au centre du site d'intérêt. Les données 360° permettent une reconstruction 3D qui est envoyée aux utilisateurs distants afin qu'ils se déplacent librement. Les utilisateurs sur site sont équipés de casques de réalité mixte pour voir où sont les utilisateurs distants sur le site. Ces derniers sont représentés sous forme d'avatar 3D (ici en bleu).

de se déplacer librement ou d'être avec d'autres utilisateurs distants simultanément. Une solution plus satisfaisante pour réunir plusieurs utilisateurs distants dans la scène consiste à intégrer une représentation 3D en plus de la vidéo. L'introduction de la 3D offre aux utilisateurs distants la liberté de se déplacer et d'occuper chacun des positions distinctes dans l'espace. (Teo *et al.*, 2019) proposent alors de combiner un flux vidéo 360° avec une reconstruction 3D du lieu. Dans ce système, un utilisateur peut basculer d'un flux 360° en direct, filmé à la première personne, à une reconstruction 3D dans laquelle il peut se déplacer librement. (Young *et al.*, 2020) adoptent aussi ce principe mais en se basant sur des caméras 360° et capteurs de profondeurs directement embarqués sur les utilisateurs. Cependant, ces systèmes se basent sur des méthodes de reconstruction 3D qui supposent que l'environnement est statique et ne sont donc pas adéquats pour des environnements dynamiques.

Notre système de télé-immersion est alors fondé sur une caméra 360° statique reconstruisant un environnement dynamique en 3D pour immerger plusieurs utilisateurs distants. Ce système devra supporter des environnements complexes comme des environnements intérieurs encombrés avec une grande variété, et pourra être le support de plusieurs cas d'usage. Il pourra servir de support de réunion mixte, par exemple en plaçant la caméra 360° dans une salle autour de laquelle sont rassemblés les participants

sur site, avec des participants à distance qui rejoignent la réunion via des avatars 3D. Il pourra aussi servir pour l'assistance à distance, par exemple pour qu'un technicien puisse être assisté par un expert distant pour l'aider à la réparation d'un appareil juste en posant une caméra 360° dans son espace de travail. Malgré le grand potentiel d'applications, nous nous sommes concentrés sur le cas d'usage de l'enseignement en ligne. L'objectif est que l'enseignant en classe pose une caméra 360° pour que des étudiants à distance suivent le cours comme s'ils étaient physiquement présents. Pour réaliser ce système, la difficulté principale réside dans l'obtention d'une reconstruction 3D d'un lieu dynamique avec une seule caméra 360° statique. Nous avons progressivement proposé des versions du système de télé-immersion afin de se rapprocher du système idéal. La première version est développée au chapitre 3, améliorée au chapitre 4 et prend sa forme définitive au chapitre 5.

1.4 Questions et Hypothèses de Recherche

Les systèmes de télé-immersion d'aujourd'hui, basés sur les caméras 360°, ont tous un inconvénient majeur. Soit tous les utilisateurs distants partagent la vue de la caméra 360°, et dans ce cas ils ne peuvent pas avoir d'avatar individuel car ils sont tous à la même position. Soit plusieurs caméras 360° sont requises, avec une caméra par utilisateur distant. La question suivante est alors posée :

> *Q1 : Est-il possible de représenter les avatars de plusieurs utilisateurs avec une unique caméra 360° ?*

La question revient à se demander si un utilisateur distant, immergé dans les données de la caméra 360°, est en capacité de voir l'avatar d'un autre utilisateur distant.

Comme nous désirons aussi que notre système permette à des utilisateurs distants de se déplacer librement, nous posons aussi la question :

> *Q2 : Est-il possible de se déplacer librement sur un site capturé uniquement avec une caméra 360° ?*

Enfin, l'approche courante pour la télé-immersion 360° consistant à diffuser à un utilisateur distant ce que la caméra est en train de filmer, nous voulons savoir si cette liberté de déplacement dans la vue 360° est souhaitée par les utilisateurs :

> *Q3 : Est-ce que les utilisateurs préfèrent la vue 360° avec ou sans navigation libre ?*

Nous formulons des hypothèses positives pour les trois questions de recherche. Nous pensons d'abord qu'il existe un moyen d'ajouter plusieurs avatars 3D dans une image 360°.

> *H1 : Il est possible de représenter les avatars de plusieurs utilisateurs dans une vue 360°.*

Nous supposons aussi qu'il existe un moyen de combiner un flux 360° avec des informations 3D, et que l'introduction de la 3D permet naturellement d'avoir les avatars de plusieurs utilisateurs dans la scène :

> *H2 : Il est possible de se déplacer sur un site capturé uniquement par une caméra 360°.*

L'information 3D avec la 360° permettant d'avoir le rendu du site à d'autres positions que celle de la caméra, les utilisateurs distants pourront se déplacer librement.

Enfin, nous supposons que la liberté de navigation va être appréciée par les utilisateurs :

> *H3 : Les utilisateurs préfèrent la vue 360° avec navigation libre.*

Bien que nous anticipions une 3D qui ne donnera pas un rendu aussi agréable que le flux 360°, la possibilité de se déplacer dans la vue 360° offre à l'utilisateur une interaction qui le fera se sentir plus présent sur le lieu distant. Se sentant moins passif avec la navigation libre, nous supposons que l'utilisateur préfère cette configuration.

1.5 Organisation de la Thèse

La mise au point d'un système de télé-immersion 3D, multi-utilisateurs, temps réel et nomade est centrale dans ce manuscrit. Les chapitres successifs iront globalement dans le sens de la réalisation de cet objectif en présentant les différentes briques améliorant de manière incrémentale le système. Le cœur du développement de ce système est précédé d'une présentation du cadre théorique justifiant nos choix et suivi d'une évaluation. Le diagramme figure 1.4 décrit cette structure du manuscrit. Le chapitre 1 discute de l'intérêt de la télé-immersion, en particulier pour l'enseignement à distance. Le chapitre 2 pose un nouveau cadre théorique de la télé-immersion. Les notions introduites seront employées tout au long du manuscrit. Le chapitre 3 détaille ce qu'est une caméra 360° et propose une première approche de télé-immersion avec cet appareil. Les images 360° étant des données limitées pour atteindre un système de télé-immersion 3D, le chapitre 4 propose une solution pour obtenir une représentation 3D avec une image 360°. Une difficulté est d'obtenir cette représentation 3D 360° pour un environnement dynamique en temps réel. Le chapitre 5 développe alors sur l'obtention de cette représentation 3D en temps réel. La supposition que cette représentation 3D 360° est adaptée pour la télé-immersion est évaluée au chapitre 6. Enfin, le chapitre 7 conclut et ouvre des perspectives pour la télé-immersion.

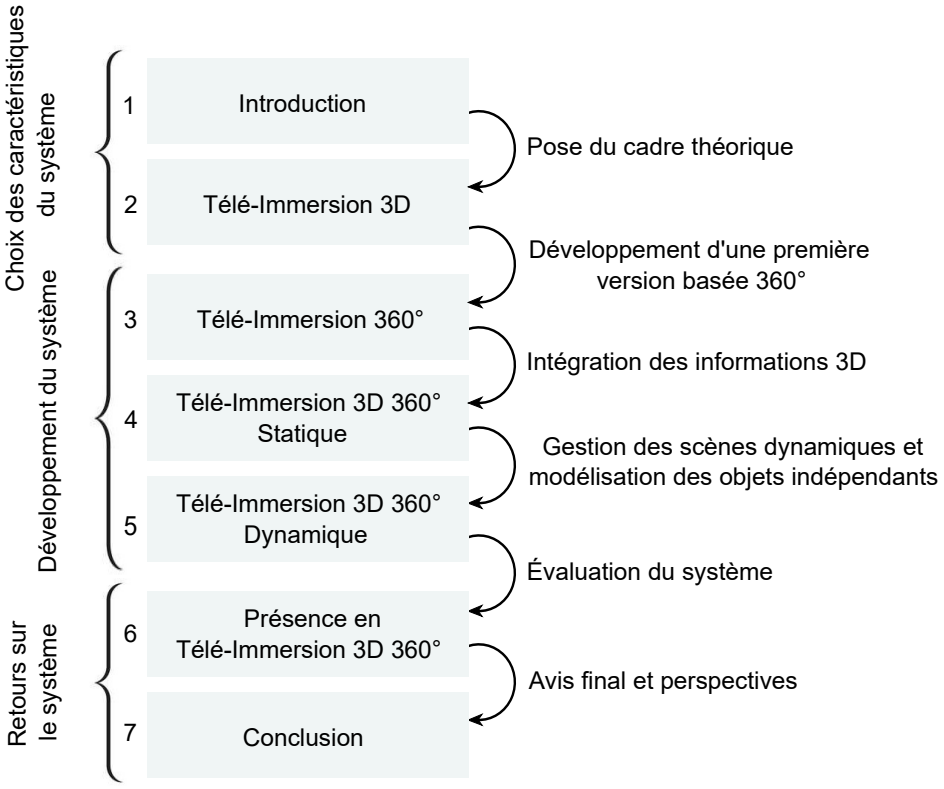


FIGURE 1.4 – Organisation du manuscrit.

1.6 Conclusion

Cette introduction donne une première esquisse de ce qu'est la télé-immersion et dépeint le système idéal que nous désirons réaliser. Celui-ci permet à des utilisateurs locaux de se réunir avec des utilisateurs distants à l'aide d'une caméra 360° et de casques de réalité étendue. Dans notre scénario, les utilisateurs sur un site d'intérêt amènent auprès d'eux des utilisateurs distants afin de collaborer comme s'ils étaient physiquement tous ensemble. Les utilisateurs sur site utilisent une caméra 360° pour effectuer une reconstruction 3D du lieu et les utilisateurs distants sont immergés dans cette reconstruction 3D 360° avec un dispositif de réalité virtuelle comme un casque de réalité virtuelle. L'apport de la représentation 3D 360° par rapport à un simple flux 360° réside dans sa capacité à permettre aux utilisateurs distants de se déplacer librement sur le site d'intérêt. Les utilisateurs sur site visualisent les utilisateurs distants à leurs positions dans la reconstruction 3D à l'aide de casques de réalité mixte. L'usage d'une unique caméra 360° avec des casques de réalité étendue permet un système nomade où la télé-immersion sur un nouveau site est possible simplement en déplaçant la caméra.

Dans les prochains chapitres, nous aborderons les étapes nécessaires à la mise en

place d'un tel système, accompagné de cas d'usage orientés pour l'enseignement à distance. Le chapitre suivant établit notre cadre de travail pour la télé-immersion.

Chapitre 2

Télé-Immersion 3D

L'introduction a mis en avant l'intérêt de la télé-immersion comme plateforme d'enseignement à distance, et le système idéal auquel nous souhaitons aboutir. Dans ce chapitre, nous introduisons formellement la télé-immersion 3D, en particulier les champs disciplinaires et les systèmes proposés dans l'état de l'art. L'étude des systèmes de télé-immersion et de leurs propriétés nous permet de proposer un nouveau cadre théorique permettant de les classer en différentes catégories. Les notions exposées vont permettre de confronter notre système de télé-immersion imaginé à la réalité et de le comparer aux systèmes avec des caractéristiques similaires comme l'acquisition avec une caméra 360° et l'immersion par réalité étendue.

2.1 À Propos de la Télé-Immersion

Le terme *télé-immersion* apparaît pour la première fois en 1996 dans le titre d'un séminaire organisé par l'Université de l'Illinois à Chicago (Leigh *et al.*, 1997). Composé du grec *télé* pour distance et du latin *immergere* pour plonger, le terme aspire à rapprocher les chercheurs en informatique distribuée, en collaboration et en réalité virtuelle. L'ambition est de permettre à des utilisateurs situés dans des lieux différents de collaborer dans un environnement partagé comme s'ils se trouvaient au même endroit (Mulligan et Daniilidis, 2001; DeFanti *et al.*, 1998). La télé-immersion se fonde alors sur les technologies de réalité virtuelle pour immerger l'utilisateur sur le lieu de collaboration. Mais, dans le contexte de la réalité virtuelle, les utilisateurs sont immergés dans un environnement virtuel généré par ordinateur (Fuchs, 2018). Or le contexte de la télé-immersion est plus général, les utilisateurs sont immergés dans un environnement distant qui peut être virtuel comme une salle modélisée en 3D, ou réel comme une salle avec une existence physique dont la représentation est retransmise par l'intermédiaire de capteurs (visuels, sonores...) (Held et Durlach, 1992). La différence est qu'en réalité virtuelle le monde est simulé (connaissance complète du modèle et des

paramètres) alors qu'en télé-immersion le monde est plutôt mesuré ou évalué (connaissance partielle à partir de données réelles). De manière plus générale, la télé-immersion profite des technologies de réalité étendue *XR*, aussi appelée *X-R* pour *X-Reality* (*X* pouvant être *virtual*, *mixed* ou *augmented*) ou réalité croisée (Rauschnabel *et al.*, 2022; Fast-Berglund *et al.*, 2018; Mann *et al.*, 2018)). La réalité étendue englobe les technologies de réalité virtuelle, réalité mixte et réalité augmentée. Certains auteurs préfèrent souligner l'aspect collaboratif de la télé-immersion en utilisant les termes Environnement Virtuel Collaboratif (*Collaborative Virtual Environment*, *CVE*), Environnement Virtuel Distribué (*Distributed Virtual Environment*, *DVE*) ou télécollaboration. Une autre dimension essentielle de la télé-immersion est le besoin de temps réel. En effet, pour avoir de l'interaction avec un élément d'un environnement distant, l'utilisateur doit avoir un retour quasi instantané de son action sur l'élément. Cet élément peut être un objet à manipuler ou une personne avec qui communiquer (sans temps réel pas de communication synchrone). La télé-immersion intègre aussi naturellement le besoin de téléopération pour pouvoir agir sur le site distant. La téléopération couvre l'ensemble des techniques permettant à un opérateur humain de se déplacer, de percevoir et de manipuler mécaniquement des objets à distance à travers des capteurs et des effecteurs (Sheridan, 1995). La téléopération est alors une composante de la télé-immersion dans son aspect d'interaction avec un environnement réel distant. Dans un système de téléopération, l'interaction avec le lieu distant est une interaction physique, il sert à réaliser une tâche mécanique. Dans un système de télé-immersion, l'interaction avec le lieu distant peut aussi être une interaction physique, par exemple si l'utilisateur contrôle un robot (sous-section 2.4.2, sous-section 2.4.3), mais nécessite en plus la communication avec les utilisateurs du site distant. Un autre concept souvent invoqué dans le contexte de la télé-immersion est celui de téléprésence. Un système de téléprésence est un système dans lequel un utilisateur se sent dans un environnement distant comme s'il y était réellement. Les systèmes de téléprésence avec un aspect collaboratif désignent généralement les mêmes systèmes que ceux de télé-immersion. La finalité d'un système de télé-immersion étant de faire collaborer des utilisateurs, le travail coopératif assisté par ordinateur est aussi être sollicité. La télé-immersion et les différents systèmes reposent alors sur une approche pluridisciplinaire, à l'intersection entre la téléprésence et le travail coopératif assisté par ordinateur.

2.1.1 Téléprésence

Le terme téléprésence est utilisé pour la première fois par Minsky en 1980 pour décrire son concept de système de téléopération (Minsky, 1980). Littéralement présence à distance, la téléprésence renvoie au phénomène selon lequel un opérateur développe le sentiment d'être présent dans un lieu distant à travers l'interaction avec le système de

téléopération (Ijsselsteijn, 2005). Les actions de l'utilisateur sur le lieu distant et le retour perceptuel qu'il reçoit résulte en un sentiment d'être physiquement sur ce lieu. Un autre terme utilisé de manière équivalente est télé-existence. Proposée indépendamment en 1980 par Tachi, la télé-existence se distingue en considérant le cas où l'utilisateur se sent présent dans un robot autonome, au contraire de la téléprésence qui se restreint strictement aux robots téléopérés (Tachi, 2019). Plusieurs facteurs sont identifiés comme augmentant la téléprésence, par exemple la transparence des dispositifs (pas de stimuli artificiels), la cohérence des informations entre les modalités ou la corrélation des actions du robot sur site avec les mouvements de l'opérateur (Held et Durlach, 1992). Le premier journal sur la téléprésence est finalement inauguré en 1992 pour étudier le phénomène (Tjostheim *et al.*, 2019). Cependant, des supports audiovisuels tentaient déjà de substituer l'environnement réel d'un spectateur par un nouvel environnement avant la téléopération et l'arrivée du terme téléprésence. Le sentiment d'être présent sur un lieu distant était expérimenté par les spectateurs au milieu du vingtième siècle avec le Cinerama et le Sensorama, ancêtres de la réalité virtuelle cinématique, ou même au dix-huitième siècle avec le panorama (Ijsselsteijn, 2005). L'idée de téléprésence s'est donc répandue dans plusieurs communautés comme l'ingénierie, la psychologie ou les sciences sociales. Le terme téléprésence a été alors remplacé progressivement par celui de présence pour prendre un sens plus large que celui en téléopération (Bourdon, 2023). La présence devient le sentiment d'être physiquement présent dans un environnement, et il est proposé de conserver la téléprésence pour le contexte de la téléopération (Sheridan, 1992). Les termes téléprésence, présence distante ou présence actuelle sont utilisés pour définir la présence dans un environnement réel médiatisé, tandis que la présence virtuelle est utilisée pour définir la présence dans un environnement virtuel simulé par ordinateur (Khenak *et al.*, 2020; Sheridan, 1992).

Parmi les chercheurs travaillant sur la présence au sens large, Bazin est reconnu comme l'un des premiers à avoir exploré le concept dans ses travaux sur le cinéma (Bazin, 1976). Il note en 1951 que pour être en présence d'une personne, il faut qu'elle existe en même temps que nous et qu'elle soit à portée de nos sens (Ijsselsteijn, 2005). Cette remarque introduit une dimension collective à la téléprésence. Les notions de présence spatiale (ou physique) et présence sociale sont adoptées pour distinguer respectivement le sentiment d'être physiquement sur un lieu et le sentiment d'être en compagnie d'une personne (virtuelle ou à distance) et de pouvoir interagir avec elle. Le sentiment simultané de présence spatiale et sociale, c'est-à-dire d'être sur un même lieu avec une personne, est appelé coprésence. Des exemples de médias avec ces différentes présences sont donnés figure 2.1. Les systèmes de téléprésence collaboratifs sont alors les systèmes de téléprésence multi-utilisateurs qui suscitent la coprésence. La collaboration en réalité étendue (Schäfer *et al.*, 2023) est incluse naturellement parmi ces systèmes en

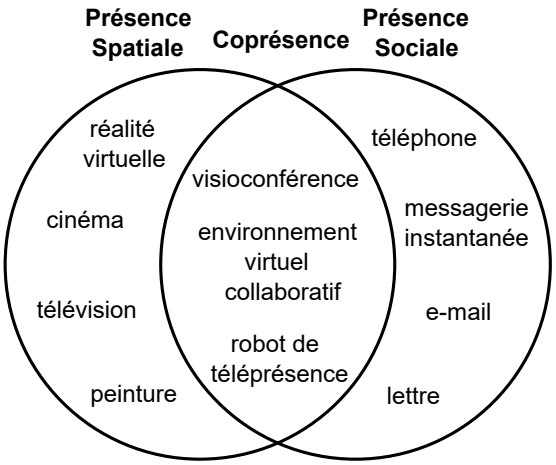


FIGURE 2.1 – Présence spatiale, présence sociale et coprésence, adapté de (Ijsselsteijn, 2005).

plus de la télé-opération. Aujourd'hui, même s'il est remplacé dans la littérature sur la réalité étendue au profit de présence, le terme téléprésence continue de définir les systèmes qui permettent à un utilisateur de se sentir accompagné dans un environnement distant ou de ressentir la présence d'un utilisateur distant dans son environnement. Enfin, une hypothèse courante suggère que l'augmentation de la téléprésence entraînerait une amélioration des performances (Pepper et Hightower, 1984). Si cette hypothèse est vraie, l'augmentation du sentiment de téléprésence des utilisateurs d'un système de télé-immersion induit une meilleure coopération entre eux.

2.1.2 Travail Coopératif Assisté par Ordinateur

Le Travail Coopératif Assisté par Ordinateur (TCAO), ou *Computer-Supported Cooperative Work* (*CSCW*), est la discipline qui étudie la manière dont les personnes travaillent ensemble par le biais des technologies de l'information et de la communication (Olson et Olson, 2003). Le domaine couvre les systèmes où les personnes collaborent dans la même pièce ou éloignées, qu'elles travaillent en même temps ou de manière désynchronisée. L'étude de l'interaction entre les personnes et les systèmes informatiques étant centrale, le TCAO s'appuie en grande partie sur les principes et les méthodes de l'Interaction Homme-Machine (IHM). L'IHM (*Human-Computer Interface*, *HCI*) s'intéresse à l'échange d'informations entre un utilisateur et un ordinateur lors d'une séquence pour contrôler l'ordinateur (point de vue de l'utilisateur) ou informer l'utilisateur (point de vue de l'ordinateur) (Hix et Hartson, 1993). L'un des principaux objectifs est de rendre l'échange d'informations plus ergonomique pour l'utilisateur. Historiquement, les travaux en IHM se concentrent sur la conception d'interfaces entre un utilisateur individuel et le système informatique. La volonté de passer de la dyade

TABLE 2.1 – Matrice Espace-Temps de (Johansen, 1988)

		Temps	
		Même	Différents
Espace	Même	**Interaction face à face** salle de décision, logiciel collaboratif à écran unique, table partagée, écran mural, matériel de salle...	**Tâche en continu** salle d'équipe, affichage public, logiciel de travail posté, gestion de projet...
	Différents	**Interaction à distance** visioconférence, messagerie instantanée, environnement virtuel collaboratif, écran partagé, éditeur multi-utilisateurs...	**Communication et coordination** e-mail, tableau d'affichage, blog, conférence asynchrone, calendrier de groupe, flux opérationnel, gestion des versions, wiki...

homme-machine à un groupe de personnes est une des raisons de l'apparition du TCAO (Bannon, 1992). Les fondements du TCAO remontent à 1945 où (Bush, 1945) initie des idées sur les technologies pour le partage d'informations (Bullinger-Hoffmann *et al.*, 2021). Ces idées inspirent Engelbart et son équipe qui lancent en 1962 des recherches sur le développement de technologies pour permettre aux individus de mieux structurer, partager et exploiter l'information (Engelbart et English, 1968). Les résultats des recherches de Engelbart sont présentés en 1968 à San Francisco lors de la *Mother of All Demos* qui introduit un grand nombre d'éléments d'interfaces informatiques modernes comme les fenêtres, la souris ou la visioconférence. Enfin, la conférence organisée en 1984 sur le thème du travail coopératif assisté par ordinateur marque la naissance de la discipline comme domaine de recherche.

Une des manières d'organiser les systèmes de TCAO est avec la matrice Espace-Temps (tableau 2.1). Introduite en 1988 par Johansen, elle divise les systèmes en quatre catégories selon qu'ils partagent ou non les mêmes lieux ou les mêmes temporalités (Johansen, 1988). Pour la dimension spatiale, la collaboration est colocalisée si les utilisateurs collaborent sur le même lieu géographique, sinon elle est distribuée. Pour la dimension temporelle, la collaboration est synchrone si les utilisateurs collaborent en

même temps, sinon elle est asynchrone. La télé-immersion se concentre alors sur la conception de systèmes collaboratifs synchrones distribués qui donnent l'illusion que les utilisateurs sont colocalisés. Le rôle de la téléprésence est crucial pour instaurer cette illusion.

Un aspect clé du TCAO dans le contexte de la télé-immersion réside dans l'analyse de la capacité d'un système à faciliter la communication entre deux utilisateurs, notamment lorsqu'ils sont à distance. La manière de représenter un utilisateur dans un système de TCAO influence la communication, par exemple des utilisateurs représentés par des avatars ont une communication non-verbale similaire à celle d'une discussion en face à face (Smith et Neff, 2018). La communication non-verbale, c'est-à-dire la communication qui ne repose pas sur la parole comme les comportements faciaux, les regards, les comportements gestuels ou les comportements spatiaux (Gruber et Kaplan-Rakowski, 2022; Maloney *et al.*, 2020), dépend donc de cette représentation. Celle-ci joue un rôle important sur l'attention des participants, la mémoire, l'efficacité de la communication ou la confiance (Jing *et al.*, 2021; Anjos *et al.*, 2019; Bohannon *et al.*, 2013). Un système de télé-immersion doit prendre en compte la représentation des utilisateurs en fonction des objectifs de collaboration et communication.

2.2 Théorie de la Télé-Immersion

L'étude des disciplines concernées par la télé-immersion permet d'appréhender les intentions des chercheurs impliqués. Nous allons désormais introduire les notions importantes pour distinguer les systèmes de télé-immersion. Nous proposons d'abord de définir la télé-immersion comme la réunion d'utilisateurs, colocalisés ou à distance, dans un lieu commun. Immergés dans un espace partagé, la télé-immersion vise à faire interagir de manière synchrone différents utilisateurs, comme s'ils étaient physiquement au même endroit quand ils sont géographiquement éloignés. Un système permettant d'accomplir cette tâche sera qualifié de système de télé-immersion. On parle spécifiquement de télé-immersion 3D lorsque la représentation de l'environnement dans lequel sont réunis les utilisateurs est en 3D. À notre connaissance, seul (Ohl, 2018) pose un cadre de travail pour la télé-immersion. Nous pensons que ce cadre introduit les notions essentiels de la télé-immersion mais ne présente pas les caractéristiques permettant de différencier ou de confondre les systèmes. Les sections suivantes exposent notre proposition de cadre de théorique exposant les caractéristiques des systèmes de télé-immersion. Une classification des systèmes rencontrés, se basant sur ces caractéristiques, est proposée (annexe A). Cette classification des systèmes dépend du dispositif d'acquisition utilisé (section 2.3) ainsi que la manière de représenter les utilisateurs distants (section 2.4).

2.2.1 Lieu

Dans notre conception de la télé-immersion, l'élément de base est le lieu où se situe la collaboration. Un lieu est un emplacement dans le monde réel ou dans un monde virtuel. On parle de lieu physique quand le lieu est dans le monde réel, comme un stade ou un bureau, et de lieu virtuel quand il est modélisé par ordinateur. Néanmoins, cette distinction entre réel et virtuel est ambiguë car un lieu virtuel peut être une reproduction d'un lieu existant réellement ou un lieu sans existence réelle intégrant des éléments réels. Cette confusion est aussi renforcée avec la généralisation des jumeaux numériques (Enders et Hossbach, 2019) représentant des objets virtuels d'objets connectés à des copies physiques, mélangeant alors des mondes virtuels avec des données réelles. On parle spécifiquement de jumeau numérique (*digital twin*) quand la connexion entre l'objet physique et la représentation virtuelle est bidirectionnelle (les deux s'influencent mutuellement) et d'ombre numérique quand la connexion est unidirectionnelle, de l'objet physiquement à la représentation virtuelle (seulement l'objet physique influence la représentation virtuelle). Le problème est alors de trancher si un lieu virtuel avec une ombre numérique est toujours un lieu virtuel. Pour différencier les lieux physiques des lieux virtuels, nous proposons de qualifier le lieu de physique uniquement si le lieu est réel et que les données sur ce lieu sont obtenues en temps réel. Sous cette condition, un emplacement réel reconstruit par photogrammétrie à partir d'une vidéo hors-ligne est alors un lieu virtuel. Un utilisateur télé-immergé sur un lieu physique est donc transporté sur un lieu géographiquement éloigné, avec un délai suffisamment faible pour considérer que l'utilisateur et le lieu sont dans la même temporalité. Bien que les données d'un lieu puissent être de différentes natures (sonores, haptiques...), nous nous concentrons tout le long exclusivement sur les données qui permettent d'obtenir des représentations visuelles. Ces informations sur l'apparence visuelle du lieu peuvent être capturées avec une simple caméra ou avec des appareils comme un LiDAR (Raj *et al.*, 2020) ou une caméra de profondeur (Zollhöfer *et al.*, 2018) pour obtenir la géométrie. Le lieu, physique ou virtuel, où les participants sont rassemblés est appelé lieu commun.

2.2.2 Scène

À partir du lieu, on construit la notion de scène. Une scène est la représentation numérique des éléments d'intérêt d'un lieu. C'est à travers une scène que l'utilisateur est immergé sur le lieu (figure 2.2). Les éléments d'intérêt d'un lieu peuvent être les utilisateurs présents sur place ou les objets à partager, et les représentations peuvent être aussi diverses que des images, des maillages, ou des primitives géométriques. Il est alors possible d'avoir plusieurs scènes pour un même lieu en utilisant une représentation différente ou en ayant différents éléments d'intérêt. La scène dépend donc directement du dispositif d'acquisition utilisé pour capturer le lieu. La mobilité permise sur le lieu

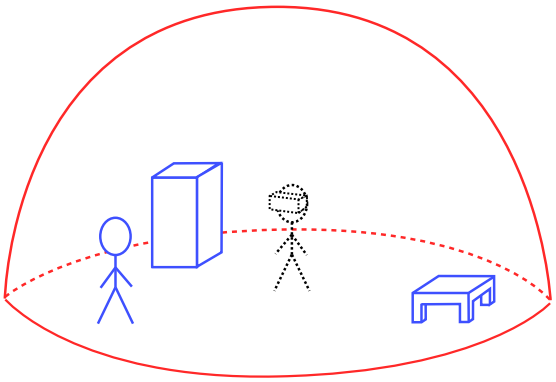


FIGURE 2.2 – Scène de télé-immersion. L'utilisateur est immergé sur le lieu distant à travers la scène. L'environnement de la scène est en rouge tandis que les objets de la scène sont en bleu.

dépend aussi directement de la scène. Par exemple, si la scène est une image 360°, un utilisateur distant n'a accès qu'à la position de la caméra 360°, là où si la scène est un maillage, un utilisateur distant peut se déplacer librement dans le lieu. La scène peut être aussi composée d'informations non-visuelles. Par exemple, une ombre numérique d'un élément physique réel sur le lieu distant peut animer son jumeau virtuel sur le lieu de l'utilisateur avec des données qui ne sont pas visuelles (Jones *et al.*, 2020b). On parle aussi de scène dynamique quand la représentation des éléments d'intérêt peut varier en fonction du temps. La difficulté des scènes dynamiques, contrairement aux scènes statiques, est que l'on doit connaître la scène à chaque instant pour avoir en temps réel l'état des éléments d'intérêt. Enfin, un système est un système de télé-immersion 3D si les éléments qui composent la scène permettent d'avoir une représentation 3D du lieu (maillage, nuage de points...).

Dans notre cadre de travail, l'opération principale de la télé-immersion est l'extension immersive (Ohl, 2018). Il s'agit de la création d'un lien immersif d'un lieu à un autre lieu distant à travers une scène, permettant à un utilisateur dans un lieu de s'étendre à un autre lieu. Ce processus se décompose en deux étapes principales. La première étape, l'extraction de scène, consiste à créer la représentation numérique du lieu distant. La deuxième étape, l'inclusion de scène consiste à présenter la scène distante dans le lieu local. Une étape intermédiaire peut être considérée, la déformation de scène, qui consiste à déformer spatialement la scène pour mettre les éléments à l'échelle du lieu local. On peut cependant voir cette étape comme faisant partie de l'extraction ou de l'inclusion de scène. Ces étapes d'extraction et d'inclusion sont exécutées respectivement à l'aide de dispositifs d'acquisition et d'immersion, aussi bien sur les lieux distants pour avoir la scène dans laquelle l'utilisateur va être immergé, mais aussi sur le lieu même de l'utilisateur afin d'avoir la scène dans laquelle les utilisateurs distants vont être immergés (figure 2.3).

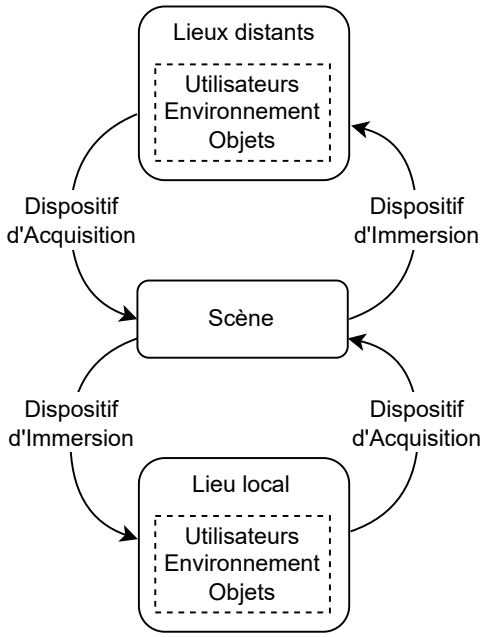
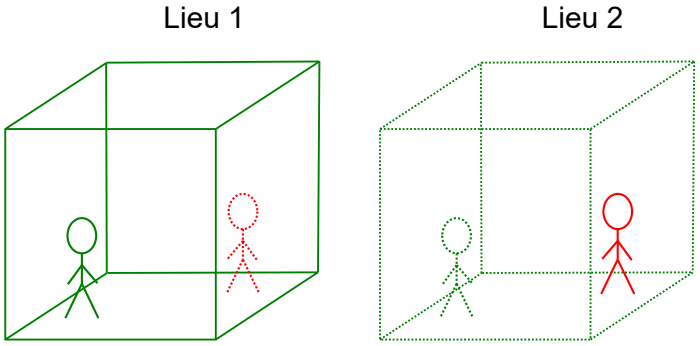


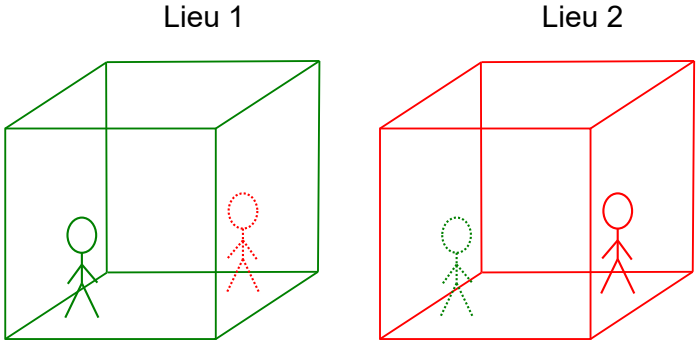
FIGURE 2.3 – Extraction et inclusion de scène.

2.2.3 Symétrie

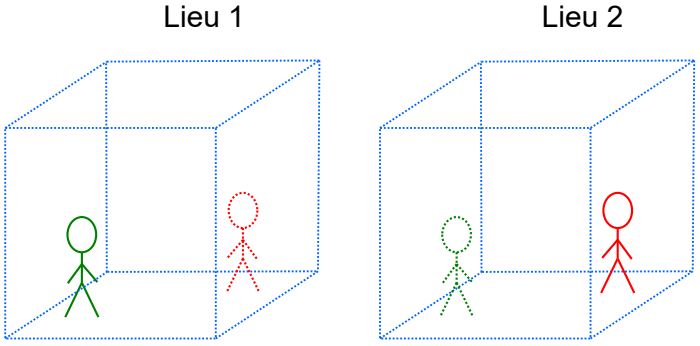
Comme relevé dans (Ohl, 2018), certains systèmes de télé-immersion sont décrits comme symétriques ou asymétriques (Rhee *et al.*, 2020; Otsuka, 2016; Nagendran *et al.*, 2015; Steed *et al.*, 2012). L'étude de l'état de l'art montre que la source de cette asymétrie dans les systèmes est le lieu commun. Si le lieu commun est un lieu physique, alors un utilisateur présent sur ce lieu va avoir un rôle différent d'un utilisateur distant. On observe alors deux catégories d'utilisateurs dans un système de télé-immersion : l'utilisateur local ou hôte, qui accueille au sein de son lieu des utilisateurs, et l'utilisateur distant, aussi appelé visiteur ou voyageur, qui est transporté dans un nouveau lieu. Cette différence est similaire à celle entre un utilisateur de réalité augmentée et de virtualité augmentée dans le continuum réalité-virtualité de Milgram (Milgram *et al.*, 1995). Dans la première, l'utilisateur perçoit le monde réel augmenté par des éléments virtuels. Dans l'autre, l'utilisateur perçoit un monde virtuel augmenté par des éléments réels. En accord avec cette perspective, (Ohl, 2018) considère chaque utilisateur comme un hôte et un visiteur en même temps à un certain degré, en fonction du nombre d'éléments de son lieu qui sont présents sur le lieu commun, le nombre d'extensions immersives. Pour notre cadre, nous ne suivons pas cette vision de continuum entre hôte et visiteur, mais une vision discrète. Certains auteurs ont proposé une séparation entre la local-présence et la téléprésence (Rauschnabel *et al.*, 2022), la première propre à la réalité augmentée, l'autre à la réalité virtuelle. En suivant cette logique, nous posons aussi qu'un utilisateur se sent de manière exclusive présent sur son lieu physique ou présent sur le lieu distant.



(a) Télé-immersion hôte-visiteur



(b) Télé-immersion hôte-hôte



(c) Télé-immersion visiteur-visiteur

FIGURE 2.4 – Catégories des systèmes de télé-immersion. Chaque lieu (représenté par un cube) abrite un utilisateur. Les différentes couleurs représentent les éléments de lieux différents. Les éléments en ligne continue représentent les éléments physiquement présents sur le lieu, ceux en ligne pointillée représentent les éléments virtuels ajoutés par la télé-immersion. (a) : L'hôte (gauche) accueille dans son lieu le visiteur (droite) qui est immergé sur le lieu hôte. (b) Chaque utilisateur est un hôte, chacun accueille dans son lieu l'utilisateur distant. (c) Chaque utilisateur est un visiteur, ils sont tous immergés sur un lieu tiers (physique ou virtuel).

Nous avons alors deux rôles d'utilisateurs distincts, hôte s'il se sent présent dans son lieu physique, visiteur s'il se sent présent sur le lieu distant. Par définition, un lieu virtuel ne peut pas contenir d'hôtes mais seulement des visiteurs. Nous identifions alors trois types de télé-immersion : hôte-visiteur impliquant des hôtes et des visiteurs, hôte-hôte impliquant uniquement des hôtes et visiteur-visiteur impliquant uniquement des visiteurs (figure 2.4). Ainsi, nous proposons de définir un système comme asymétrique s'il implique au moins un hôte et un visiteur. C'est-à-dire qu'un système asymétrique fait collaborer un utilisateur qui se sent présent sur un lieu qui n'est pas le sien avec un utilisateur qui est sur son lieu physique (Rhee *et al.*, 2020; Jones *et al.*, 2020a; Steed *et al.*, 2012). Un système hôte-visiteur est donc naturellement asymétrique. À l'inverse, nous posons qu'un système symétrique n'implique que des hôtes ou que des visiteurs. Un système hôte-hôte où tous les utilisateurs sont des hôtes est symétrique. Dans ces systèmes, chaque utilisateur amène auprès de lui un ou plusieurs utilisateurs géographiquement éloignés (Zhang *et al.*, 2022b; Lawrence *et al.*, 2021; Orts-Escolano *et al.*, 2016; Pejsa *et al.*, 2016). Un système visiteur-visiteur où tous les utilisateurs sont des hôtes est aussi symétrique. Dans ces systèmes, chaque utilisateur est transporté dans un autre lieu, virtuel ou réel, sans utilisateurs locaux. Dans ce cas, le lieu commun est généralement un lieu virtuel.

Une subtilité dans notre théorie concernant le lieu commun est qu'il peut ne pas être unique. Par exemple, avec une télé-immersion visiteur-visiteur où les utilisateurs sont réunis sur un lieu virtuel, il est possible que le lieu virtuel ne soit pas le même pour tous. Si les utilisateurs décident la décoration du lieu virtuel, certains peuvent opter pour des environnements différents des autres, le lieu commun n'est pas le même selon les préférences individuelles. C'est aussi toujours le cas pour une télé-immersion hôte-hôte. Tous les utilisateurs étant des hôtes, chacun reçoit les autres utilisateurs dans son lieu physique. Il y a donc autant de lieux communs que d'hôtes. Mais la non-unicité du lieu commun peut poser des difficultés pour la télé-immersion hôte-visiteur. Notre cadre permet qu'un tel système puisse avoir des hôtes répartis sur plusieurs lieux physiques, par exemple un hôte unique par lieux physiques. Le problème est alors de savoir vers quel lieu un visiteur est transporté. Une possibilité est qu'il soit immergé sur un lieu virtuel tiers où il est réuni avec les représentations des hôtes. Ce lieu virtuel est donc un nouveau lieu commun en plus des autres lieux physiques, et tous ont une importance équivalente. Une autre possibilité est que le visiteur soit immergé sur un lieu physique d'un hôte en particulier, délaissant alors les autres lieux physiques. Dans ce cas-là, on a autant de lieux communs que de lieux physiques, mais l'intuition est que celui où est immergé le visiteur a plus d'importance que les autres lieux communs. On a alors une asymétrie des lieux communs, où celui qui contient le plus d'utilisateurs (l'hôte et le visiteur) est plus important que les autres (avec seulement un hôte). Cette asymétrie

peut être décrite avec une valeur comptant le nombre d'utilisateurs pour chaque lieu. Néanmoins, nous pensons que cette asymétrie dans le lieu commun est superficielle comparée à l'asymétrie basée sur la différence de rôle entre hôte et visiteur. De plus, nous n'avons pas trouvé de système asymétrique concret dans ce cas de figure avec plusieurs lieux communs. Dans un système hôte-visiteur, la norme est que les hôtes sont tous sur le même lieu physique et que les visiteurs sont transportés sur ce lieu.

Pour conclure, notre définition de la symétrie est différente de celle de (Ohl, 2018) qui pose plusieurs types symétries. Une idée générale est proposée pour qualifier un système de symétrique si l'extraction et l'inclusion est la même sur tous les lieux. Il propose aussi d'analyser deux symétries sur l'extraction de la scène. La première symétrie d'extraction est qualitative, le système est symétrique si des éléments de même nature sont capturés sur chaque lieu. La seconde symétrie d'extraction est quantitative, elle considère le système symétrique si le lieu commun contient équitablement des éléments des différents lieux, c'est-à-dire un nombre d'extensions immersives égale pour chaque lieu. Avec cette définition, le système est asymétrique si le lieu commun contient plus d'éléments d'un lieu en particulier. Enfin, une dernière symétrie est la symétrie de communication. Cette symétrie considère un système comme symétrique s'il y a le même nombre d'utilisateurs sur les lieux de la télé-immersion, le système est asymétrique s'il y a plus d'utilisateurs sur un lieu que sur un autre. Nous pensons que notre définition de la symétrie sur la différence entre hôtes et visiteurs est plus pertinente car elle capture de manière plus fondamentale les différents types de télé-immersion. Avec notre définition, on a bien deux types d'utilisateurs avec deux rôles différents dans un système asymétrique, les uns accueillent, les autres sont reçus. Dans un système symétrique, tous les utilisateurs sont égaux, peu importe leurs lieux d'origine. Cette symétrie s'appuyant sur le rapport des utilisateurs au lieu commun, la manière de capturer un lieu lors de l'extraction est centrale pour déterminer la symétrie. Cependant, notre définition de symétrie n'impose pas que tous les utilisateurs utilisent le même dispositif d'extraction et d'inclusion de scène. Il existe alors des systèmes de télé-immersion asymétriques qui utilisent les mêmes dispositifs sur les différents lieux (Young *et al.*, 2020). Nous soutenons tout de même que notre définition de la symétrie implique une différence dans l'extraction et l'inclusion de la scène.

2.3 Extraction de Scène

Comme présenté, la télé-immersion repose sur l'extraction de la scène et l'inclusion de la scène. L'extraction de la scène est réalisée grâce à des dispositifs d'acquisition pour obtenir une représentation des éléments d'intérêt du lieu. Cependant, même si les données acquises peuvent ne pas être directement visuelles, comme avec une ombre numérique où les données des capteurs sont des informations pouvant décrire la forme, la

position ou le mouvement d'un objet réel, elles ont toujours vocation à être présentées visuellement à l'utilisateur. Les données d'une ombre numérique vont alors toujours être utilisées par l'utilisateur pour visualiser un objet d'intérêt, que ce soit avec une représentation virtuelle comme modèle 3D ou une représentation matérielle comme une copie physique de l'objet (voir section 2.4). La capture de ces données non-visuelles qui sont finalement rendues est largement utilisée pour représenter les avatars. En réalité virtuelle, il est commun de supposer qu'un utilisateur distant connaît une représentation 3D d'une personne (avatar reconstruit à l'avance ou avatar générique) et que les informations de suivi du casque et des manettes permettent d'animer cette représentation (Kim *et al.*, 2023; Luo *et al.*, 2023; Ma *et al.*, 2021; Jones *et al.*, 2020a; Rhee *et al.*, 2020; Tan *et al.*, 2017). Le principe est aussi utilisé avec des caméras pour animer l'orientation du regard ou la pose de la représentation d'un utilisateur (He *et al.*, 2021; Yu *et al.*, 2021; Shiro *et al.*, 2018; Otsuki *et al.*, 2017). Dans ces cas cités, la scène n'est pas une représentation directement visuelle.

Dans cette section, nous présenterons les types de dispositifs d'acquisition et les caractéristiques pertinentes pour différencier les systèmes en nous intéressant uniquement aux dispositifs capturant une représentation directement visuelle de la scène. Certains dispositifs ne sont actuellement pas utilisés pour la télé-immersion mais pourraient naturellement servir de base pour un futur système. Les méthodes pour animer des représentations connues à l'avance ne seront pas traitées, car elles sont généralement simples (par cinématique inverse). De plus, notre objectif est d'avoir un système de télé-immersion nomade où la simple pose du dispositif d'acquisition suffit à télé-immerger un utilisateur sans avoir besoin de reconstruire à l'avance des éléments du lieu.

2.3.1 Acquisition *Outside-In* et Acquisition *Inside-Out*

Il existe une grande variété dans les dispositifs d'acquisition qui sont proposés aujourd'hui : mono-caméra, multi-caméras, statique, dynamique... Cependant, nous observons qu'il existe une séparation notable dans les dispositifs d'acquisition en fonction du type d'élément qui doit être capturé : les dispositifs *outside-in* et *inside-out*. Les premiers vont capturer ce qu'on appellera des objets, tandis que les derniers vont capturer des environnements. Le choix entre ces deux catégories de dispositifs est essentiel pour la télé-immersion car il va favoriser le type d'usage du système.

Objet et Environnement

Dans la création de contenu pour la réalité virtuelle, certains auteurs ont noté une distinction dans le type d'éléments qui composent une scène en séparant les éléments champ proche et les éléments champ lointain. (Richardt *et al.*, 2020) différencient un *objet* d'une *scène* selon le dispositif d'acquisition utilisé. (Dupont de Dinechin, 2020)

distingue l'*environnement de fond* qui correspond aux éléments de contexte hors de portée, des *objets de premier plan* qui sont les éléments dans un espace proche avec lequel il est possible d'interagir. Dans le contexte de la télé-immersion, (Steed *et al.*, 2012) appellent *destination* la représentation du lieu dans laquelle l'utilisateur distant est immergé. En suivant ces observations, nous considérons que deux types d'éléments composent une scène de télé-immersion : *objet* et *environnement*. Un exemple de la composition d'une scène est donné figure 2.2. Nous proposons pour un environnement et un objet les définitions suivantes. Un environnement correspond à la surface d'un volume à l'intérieur duquel un utilisateur est placé. Ce volume représente les limites de la scène de télé-immersion, un utilisateur ne peut pas se déplacer au-delà de l'environnement. Un environnement ayant vocation à modéliser les éléments en champ lointain, l'utilisateur ne peut pas se déplacer autour d'un élément de l'environnement. Cette restriction étant d'ordre topologique (pour tourner complétement autour d'un élément, celui-ci doit appartenir à une autre composante connexe), certaines topologies sont adaptées pour représenter le volume de l'environnement. Un environnement étant généralement vu comme une représentation globale indivisible, les données utilisées pour modéliser un environnement peuvent ne pas contenir suffisamment d'informations pour connaître les bords de chaque élément individuel qui le compose. Il est alors impossible de sélectionner ou manipuler un élément particulier de l'environnement. Dans la plupart des moteurs de rendu 3D, l'environnement est modélisé par une *skybox*, c'est-à-dire texture de fond pour représenter un horizon inatteignable. À noter qu'il peut y avoir plusieurs environnements dans une scène. Si la scène est un ensemble d'images 360° avec un utilisateur se téléportant d'une image à l'autre, alors chaque image 360° représente un environnement. L'autre type d'élément qui compose une scène, un objet, consiste simplement en un élément à l'intérieur de l'environnement. Celui-ci doit être topologiquement indépendant du volume de l'environnement. Contrairement à l'environnement, un objet représente un élément en champ proche, l'utilisateur peut donc se déplacer autour d'un objet sans restrictions. Un objet étant topologiquement indépendant de l'environnement, celui-ci peut être sélectionné et manipulé individuellement. Un type d'objet particulièrement important pour la télé-immersion est l'utilisateur. En effet, un système de télé-immersion visant à rassembler des utilisateurs sur un lieu commun, une représentation des personnes doit être contenue dans la scène afin qu'ils puissent se voir et interagir entre eux. Le diagramme figure 2.5 résume la hiérarchie des éléments qui composent une scène.

Dispositif *Outside-In* et *Inside-Out*

Comme noté dans la littérature, cette distinction entre environnement et objet se justifie aussi dans les dispositifs d'acquisition utilisés pour passer d'un élément réel à une

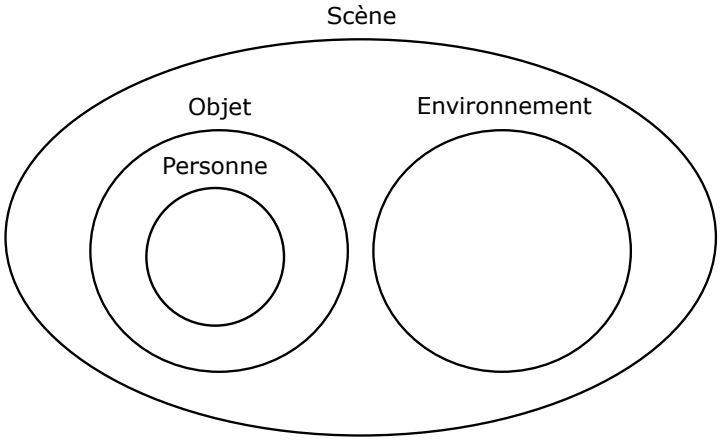


FIGURE 2.5 – Type d'éléments d'une scène de télé-immersion. Un environnement et un objet sont des éléments de nature différente, tandis qu'une personne est un objet particulier.

scène. Malgré leurs diversités, la majorité des dispositifs peut être aujourd'hui classée en deux catégories : *outside-in* et *inside-out*, ou *inward* et *outward* (Dupont de Dinechin, 2020; Richardt *et al.*, 2020; Kowalski *et al.*, 2015; Lee *et al.*, 2015). Même si ces dispositifs d'acquisition peuvent être mono-caméra, multi-caméras avec une disposition quelconque, avec une caméra dynamique (Newcombe *et al.*, 2011) ou avec un objet d'intérêt mobile (Jiang *et al.*, 2022), ils peuvent être répartis selon un modèle simple. Ce modèle consiste à approximer l'ensemble des dispositifs d'acquisition comme des dispositifs multi-caméras arrangées en cercle. La distinction entre *outside-out* et *outside-in* dépend uniquement de l'orientation des caméras du dispositif. Si celles-ci sont orientées vers l'intérieur du cercle, alors le dispositif est qualifié d'*outside-in*. À l'inverse, si les caméras sont orientées vers l'extérieur, le dispositif est qualifié d'*inside-out*. Dans un dispositif *outside-in*, l'élément d'intérêt à capturer est placé à l'intérieur du cercle pour être visible sous plusieurs points de vue. Sur les images de chacune des caméras, l'élément d'intérêt doit être segmenté pour le séparer de l'arrière-plan. Une stratégie commune dans les studios est d'utiliser un fond vert pour faciliter cette segmentation (Zins *et al.*, 2021; Collet *et al.*, 2015). Les différentes images des caméras sont combinées pour obtenir un modèle 3D de l'élément d'intérêt. Avec cette configuration, les axes optiques des caméras se recoupent près de l'élément d'intérêt, créant des chevauchements entre les images qui facilitent la reconstruction 3D. Des exemples de dispositifs d'acquisition *outside-in* sont donnés figure 2.6. Avec des prises de vues tout autour de l'élément, celui-ci va pouvoir être reconstruit sous forme de volume et va donc naturellement pouvoir représenter un objet dans une scène. D'autre part, dans un dispositif *inside-out*, l'élément d'intérêt correspond à l'ensemble du lieu, à l'extérieur du cercle, sans distinctions entre les éléments. Les scènes capturées en configuration *inside-out* sont aussi

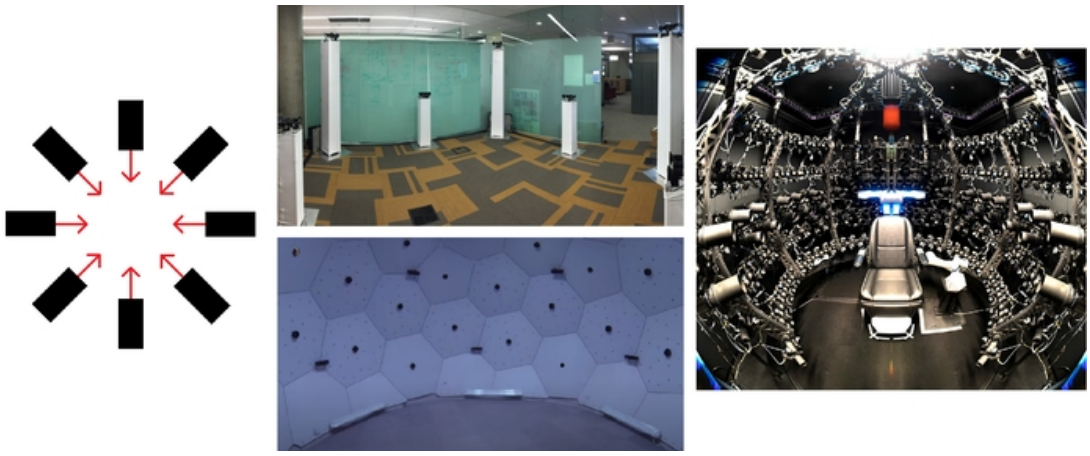


FIGURE 2.6 – Dispositifs d'acquisition *outside-in*. Gauche : Modèle d'acquisition *outside-in*. Milieu-Haut : (Orts-Escolano *et al.*, 2016). Milieu-Bas : (Joo *et al.*, 2015). Droite : (Cao *et al.*, 2022).

appellées *unbounded scene* dans le contexte des NeRFs (Barron *et al.*, 2022). Les appareils comme les LiDARs, les caméras *Light Field* (Broxton *et al.*, 2020; Overbeck *et al.*, 2018) ou les caméras 360° appartiennent à cette catégorie. Des exemples sont donnés figure 2.7. Les dispositifs *inside-out* ont un avantage sur les dispositifs *outside-in* pour notre système de télé-immersion : ils sont généralement plus facilement transportables. Le besoin de capturer l'élément d'intérêt sous plusieurs points de vue implique que la majorité des dispositifs *outside-in* sont plus encombrants (incluant des salles dédiées) et pas forcément adaptés à un usage nomade. À l'inverse, un dispositif *inside-out* est souvent conçu comme un unique appareil, ce qui rend son transport dans d'autres lieux possible. Cependant, la reconstruction 3D avec ce type d'appareils est plus compliquée qu'en configuration *outside-in*. Contrairement à la capture *outside-in*, les axes optiques des caméras ne se recoupent pas en un point. La capture est alors plus difficile, car pour un même nombre d'images, elle entraîne un nombre de chevauchements entre les images bien inférieur (Hedman *et al.*, 2016). De plus, cette configuration ne vise pas à capturer un élément individuel mais l'ensemble du lieu qui peut être encombré. Or, cette configuration consistant en une unique vue ou à des vues relativement proches, il y a besoin de gérer les occultations et les auto-occultations qui créent des régions invisibles (Li *et al.*, 2019). La capture globale de plusieurs éléments sans distinctions, et donc sans possibilités d'interactions avec un élément individuel, est plus adaptée à représenter l'environnement de la scène.

Cette séparation simple entre *outside-in* et *inside-out* en fonction de la convergence des caméras permet de classer une bonne partie des dispositifs. Cependant, il reste difficile pour certains dispositifs de trancher entre *outside-in* et *inside-out*. Par exemple, une unique caméra perspective ou un ensemble de caméras disposées en ligne

FIGURE 2.7 – Dispositifs d'acquisition *inside-out*. Gauche : Modèle d'acquisition *inside-out*. Milieu-Gauche : (Bertel *et al.*, 2020). Milieu-Droite : (Broxton *et al.*, 2020). Droite : (Overbeck *et al.*, 2018).

n'appartiennent à aucune catégorie. Pour généraliser cette classification, nous pouvons considérer les dispositifs à la limite comme *outside-in* ou *inside-out* respectivement en fonction de ce qu'ils cherchent à capturer, un élément précis du lieu d'un point de vue extérieur ou une vision globale du lieu vu de l'intérieur.

Systèmes de Télé-Immersion

Nous allons présenter quelques systèmes de télé-immersion notables de l'état de l'art. Le tableau A.1 résume les systèmes étudiés avec leurs dispositifs d'acquisition. Puisque la raison des dispositifs *outside-in* est de capturer un élément précis du lieu, ceux-ci vont être utilisés pour capturer une représentation de l'utilisateur qui va être visionnée par les autres. Ceux-ci vont alors être fréquemment utilisés pour les systèmes symétriques hôte-hôte et visiteur-visiteur. L'emploi le plus commun pour la télé-immersion est alors de placer l'utilisateur est au centre du dispositif *outside-in* pour que les autres perçoivent sa représentation 3D (Córdova-Esparza *et al.*, 2019; Gotsch *et al.*, 2018; Zioulis *et al.*, 2016; Roberts *et al.*, 2015; Beck *et al.*, 2013; Gross *et al.*, 2003). Le meilleur représentant de cette configuration est Holoportation (Orts-Escolano *et al.*, 2016) qui capture l'utilisateur en 3D avec un ensemble de caméras RGB-D autour de lui pour l'afficher aux utilisateurs distants en réalité augmentée (figure 2.13 milieu). Il est aussi possible que les systèmes que nous définissons comme symétriques implique des utilisateurs avec des dispositifs différents. Dans (Jones *et al.*, 2009), un utilisateur distant est capturé en 3D avec des caméras tout autour de lui, mais les autres utilisateurs sont capturés simplement à l'aide d'une caméra (pas de représentation 3D). Dans ce cas, on a une télé-immersion hôte-hôte où chacun à l'impression, plus ou moins crédible, que l'autre utilisateur est avec lui sur son site. De l'autre côté, les dispositifs *inside-out* capturant l'environnement vont être utilisés de préférence pour les systèmes asymétriques avec des utilisateurs qui restent sur le lieu, et les autres qui sont immergés dans l'environnement. Aujourd'hui, la majorité des dispositifs *inside-out* pour la télé-immersion sont

des caméras 360°. Cette caméra peut être statique (Rhee *et al.*, 2020; Shiro *et al.*, 2018), contrôlable à distance (Kim *et al.*, 2023; Jones *et al.*, 2020a; Heshmat *et al.*, 2018; Ogi et Fueki, 2017) ou embarquée par un hôte (Piumsomboon *et al.*, 2019; Teo *et al.*, 2019; Lee *et al.*, 2018; Kasahara et Rekimoto, 2015). Le système proposé par (Stotko *et al.*, 2019a) représente bien les systèmes *inside-out* basés sur un appareil autre qu'une caméra 360° : un utilisateur navigue à travers le lieu avec une caméra de profondeur et reconstruit progressivement le lieu en 3D. Les utilisateurs distants naviguent librement dans cette reconstruction en réalité virtuelle et peuvent demander une mise-à-jour d'une région précise si une modification y a été effectuée. Une évolution de ce système a été développée en se basant sur un robot téléopéré (figure 2.8 droite). Enfin, en plus d'être *inside-out*, ces systèmes sont souvent dans une configuration égocentrique (sous-section 2.3.2).

Comme expliqué plus haut, les dispositifs *inside-out* sont plutôt la base des systèmes de télé-immersion nomades, où des utilisateurs distants sont immergés dans un nouveau lieu simplement en déplaçant le dispositif d'acquisition. À notre connaissance, FarfetchFusion (Lee *et al.*, 2023) est le seul système de télé-immersion nomade clairement conçu avec un dispositif d'acquisition *outside-in* afin de capturer une représentation 3D du visage de l'utilisateur grâce à un appareil composé de plusieurs smartphones. Des systèmes peuvent aussi être vus comme portables car le dispositif consiste en un ensemble de caméras qu'il convient de transporter et disposer sur le nouveau lieu comme Holoportation (Orts-Escolano *et al.*, 2016). Enfin, certains systèmes portables sont difficiles à classer entre *outside-in* et *inside-out* comme HoloKinect (Siemonsma et Bell, 2022) ou la version portable de Room2Room (Pejsa *et al.*, 2016) car reposant sur une unique caméra perspective avec profondeur capturant seulement la représentation des utilisateurs. Bien que cette distinction entre dispositif *outside-in* et dispositif *outside-in* serve à justifier la différence entre objet et environnement, il est possible en pratique d'avoir des dispositifs *outside-in* qui capturent des représentations d'environnements et des dispositifs *inside-out* qui capturent des représentations d'objets. Des systèmes comme (Kolkmeier *et al.*, 2018; Maimone et Fuchs, 2012) reposent sur un ensemble de caméras Kinect réparties sur le lieu commun, plutôt propre aux systèmes *outside-in*, qui ne capturent pas un objet particulier mais l'ensemble du lieu, utilisateurs inclus. Il existe alors des systèmes de télé-immersion où un élément capturé par un dispositif *inside-out* représente un objet dans la scène. Concevoir de tels systèmes représente un défi étant donné les difficultés de la reconstruction 3D avec ce dispositif. La principale étant que l'objet est vu uniquement de face, impliquant que des parties cachées doivent potentiellement être reconstruites pour le modéliser en 3D. Mais il est possible de faire usage de représentations simplifiées. Dans (Dasari *et al.*, 2023), les utilisateurs sont réunis dans un lieu virtuel 3D où chacun est représenté par un plan 2D positionné dans l'espace 3D, sur lequel leur flux vidéo est projeté. Ce flux vidéo est créé avec une webcam fixe

devant les utilisateurs. Dans Mobileportation (Young *et al.*, 2020), les utilisateurs sont aussi représentés par un plan vidéo dans la scène 3D, chacun capturé grâce à une caméra 360°. (Deng *et al.*, 2023) capturent des utilisateurs grâce à une caméra perspective et exploitent un algorithme d'estimation de pose pour représenter les utilisateurs sous forme de plan vidéo avec du relief aux bras.

Pour conclure, la différence entre objet et environnement est essentielle pour la télé-immersion. Selon (Ohl, 2018), le lieu où se déroule la télé-immersion dépend du nombre d'éléments qui composent la scène, tel que plus les éléments d'un lieu sont présents sur le lieu commun, plus le lieu commun tend à être ce lieu. Nous rejetons cette définition du lieu commun car notre séparation entre environnement et objet nous laisse penser que les éléments d'une scène n'ont pas tous le même poids sur le lieu commun. Nous proposons que si une scène créée à partir d'un certain lieu contient un environnement, alors ce lieu est le lieu commun. Si l'environnement d'un lieu est capturé et présenté, ce lieu est le lieu commun. Nous soutenons que les objets ne jouent pas de rôle sur la définition du lieu commun. En effet, si un environnement n'est pas présenté à l'utilisateur, celui-ci continu de se sentir sur son lieu et est donc un hôte. Même en présentant un grand nombre d'objets d'autres lieux, cet utilisateur continu de se sentir dans son lieu, juste avec des objets en plus. À l'inverse, si un environnement est présenté à l'utilisateur, son lieu physique disparait pour être remplacé par le lieu distant. Dans ce cas, l'utilisateur est un visiteur. Peu importe s'il amène avec lui un grand nombre d'objets de son lieu sur le lieu commun, il se sent sur l'autre lieu. Notre définition de la symétrie se rapproche alors de l'aspect qualitatif de la symétrie d'extraction de (Ohl, 2018). Un environnement et un objet n'étant pas de même nature, on peut considérer le système comme asymétrique s'il implique ces deux types d'éléments. On remarque que lorsqu'un système est asymétrique, un dispositif *inside-out* est utilisé sur le lieu de l'hôte pour que la télé-immersion se déroule sur ce lieu. Quand il est symétrique, un dispositif *outside-in* est utilisé sur le lieu d'un hôte ou d'un visiteur pour créer les représentations visuelles des utilisateurs. Choisir un dispositif *inside-out* comme une caméra 360° est donc relativement naturel pour un système nomade et asymétrique. Enfin, dans le cas où la scène mélange des environnements de plusieurs lieux différents, nous posons que le lieu commun est un lieu virtuel. Dans ce cas, l'environnement ne correspond plus à un environnement réellement existant. En pratique, nous n'avons pas trouvé de système de télé-immersion avec cette approche.

2.3.2 Acquisition Exocentrique et Acquisition Égocentrique

Comme défini précédemment, un dispositif d'acquisition *outside-in* est placé autour de l'objet à capturer tandis qu'un dispositif d'acquisition *inside-out* est au sein même de l'environnement. Ainsi, dans une configuration *outside-in* le dispositif se trouve à

FIGURE 2.8 – Systèmes de téléprésence égocentriques. Gauche : Polly, une caméra portée à l'épaule montée sur un mécanisme contrôlable par l'opérateur à distance (Kimber *et al.*, 2015; Kratz *et al.*, 2014). Milieu : Caméra 360° fixée sur un sac à dos (Tang *et al.*, 2017). Droite : Mario, un bras robotique avec une caméra RGB-D embarquée sur une base mobile (Stotko *et al.*, 2019b).

.

l'extérieur du lieu alors que dans une configuration *inside-out* le dispositif se trouve à l'intérieur du lieu. Mais un cas particulier d'acquisition *inside-out* est de considérer que les informations sont capturées avec un appareil embarqué par un objet comme une personne. Les informations peuvent alors être considérées comme étant le point de vue subjectif de l'objet. Une distinction dans le type de scène apparaît alors en fonction de la perspective à laquelle les données sont perçues : la scène peut être à la troisième personne ou à la première personne. Nous proposons la distinction entre acquisition égocentrique et acquisition exocentrique (Young *et al.*, 2020) pour nommer cette différence. L'acquisition égocentrique consiste à capturer la scène du point de vue d'un objet à l'intérieur du lieu (*inside-out* embarquée). Un utilisateur visualisant des données égocentriques perçoit directement le lieu comme s'il était à la place de l'objet. L'acquisition exocentrique consiste à capturer la scène avec une vision extérieure à la scène (*outside-in* ou *inside-out* non-embarquée). Ces données ne sont asservies au point de vue d'aucun objet du lieu.

Systèmes de Télé-Immersion Égocentrique

La majorité des systèmes de télé-immersion reposant sur une acquisition exocentrique, nous allons présenter ici des propositions se basant sur une acquisition égocentrique. Des exemples de ces systèmes sont donnés figure 2.8. Le principe est de placer le système d'acquisition sur un hôte ou sur un robot afin que le visiteur soit immergé à leur place. L'hôte ou le robot vont pouvoir se déplacer pour se rendre à l'endroit où la collaboration doit avoir lieu. La première catégorie de système de télé-immersion égocentrique consiste donc à immerger un visiteur à la place d'un hôte sur le lieu. Pour qu'un visiteur puisse s'immerger à la place d'un hôte, ce dernier va utiliser un système

de téléprésence portable. Ce système consiste simplement en une caméra embarquée, portée généralement sur la tête, sur l'épaule ou sur le torse (Pfeil *et al.*, 2019), qui va retransmettre en direct le point de vue de l'hôte. Le premier système dans ce sens à se qualifier de système de téléprésence portable est proposé par (Drugge *et al.*, 2004), même si d'autres systèmes antérieurs peuvent être considérés comme tel (Sakata *et al.*, 2003; Tsumaki *et al.*, 2002). Le système est composé d'une caméra montée sur la tête avec un casque de réalité virtuelle pour visualiser les participants à distance. Le problème de ce système est que le visiteur, contraint par la position et l'orientation de la caméra, a peu de liberté dans le choix de son point de vue. Pour donner plus de liberté, certains auteurs ont présenté des systèmes avec un mécanisme contrôlable par le visiteur pour changer l'orientation de la caméra. Ces systèmes sont généralement portés à l'épaule comme Polly (Kimber *et al.*, 2015; Kratz *et al.*, 2014), MH-2 (Han *et al.*, 2018; Tsumaki *et al.*, 2012) ou TEROOS (Kashiwabara *et al.*, 2012). Une autre piste explorée est l'utilisation d'une caméra 360° au lieu d'une caméra perspective classique (Piumsomboon *et al.*, 2019; Young *et al.*, 2019; Lee *et al.*, 2018; Tang *et al.*, 2017; Kasahara et Rekimoto, 2015) permettant aux participants de choisir la direction dans laquelle regarder. D'autres systèmes sont également élaborés pour capturer les informations visuelles en 3D pour que le visiteur se déplace virtuellement sans imposer de déplacements physiques à l'hôte. (Teo *et al.*, 2019) ont proposé d'utiliser une caméra 360° embarquée avec une reconstruction 3D du lieu afin que le visiteur puisse voir le lieu du point de vue de l'hôte ou choisir indépendamment son propre point de vue. Cependant, la reconstruction 3D du lieu doit être réalisée a priori, ce qui limite la télé-immersion à des sites connus. JackIn Space (Komiyama *et al.*, 2017) est un système permettant de choisir entre une vue à la première personne, obtenue avec une caméra *fisheye* embarquée, et une vue à la troisième personne consistant en une reconstruction 3D du lieu à partir de caméras RGB-D disposées sur le plafond. Ce système nécessite alors une instrumentation préalable du lieu, inutilisable pour des sites inconnus. Des systèmes de téléprésence portables intégrant une caméra de profondeur ont alors été proposés, permettant ainsi de maintenir une reconstruction 3D globale du lieu, actualisée en temps réel en fonction de la position de l'hôte (Stotko *et al.*, 2019a; Gao *et al.*, 2017; Sodhi *et al.*, 2013). Un visiteur peut alors solliciter l'hôte pour reconstruire spécifiquement une région du lieu. L'idée de pouvoir passer d'une vue égocentrique à une reconstruction 3D a été reprise dans Mobileportation (Young *et al.*, 2020) en proposant de reconstruire le lieu en temps réel en combinant la caméra 360° avec une caméra de profondeur. Cependant, une majorité de ces systèmes font collaborer un ou plusieurs visiteurs avec seulement un hôte. Certains systèmes tentent alors de faire participer d'autres hôtes qui ne seraient pas équipés d'un système de téléprésence portable. Par exemple, JackIn Neck (Matsuda *et al.*, 2018) et Gutsy-Avatar (Tobita, 2017) proposent de porter un écran en plus des

caméras, afin que les autres hôtes voient le flux vidéo du visiteur. Ainsi, le système permet une communication tripartite entre le visiteur, l'hôte portant l'appareil, et un autre hôte. De manière originale, ChameleonMask (Misawa et Rekimoto, 2015) vise à ce que le visiteur prenne la place de l'hôte aux yeux des autres personnes sur site. Pour atteindre ce but, l'hôte porte un casque de réalité virtuelle sur lequel est attachée une tablette qui masque son visage et diffuse à la place le flux vidéo du visiteur.

L'acquisition égocentrique est aussi pertinente lorsque l'utilisateur peut directement contrôler le dispositif d'acquisition sur le lieu, comme avec un robot. Dans cette catégorie, l'utilisateur est immergé sur le lieu à travers un robot, qui peut être un robot de téléprésence (sous-section 2.4.2) ou un robot anthropomorphique (sous-section 2.4.3). Le visiteur téléopére le robot pour le déplacer physiquement afin de percevoir le lieu sous un nouveau point de vue (Youssef *et al.*, 2023; Kristoffersson *et al.*, 2013). Dans ce cas, la scène transmise aux visiteurs se limite souvent à une simple vidéo, perspective ou 360°, capturée depuis la position du robot. Mais les données capturées par les robots peuvent également être utilisées pour créer des scènes 3D. Par exemple, (Sumigray *et al.*, 2021) ont proposé un robot pour la téléopération, aussi équipé d'une caméra 360° mais avec des caméras de profondeur pour créer une scène 3D avec des parallaxes de mouvement sans déplacer physiquement le robot. De son côté, (Stotko *et al.*, 2019b) ont proposé d'intégrer une caméra de profondeur sur un robot mobile afin d'explorer le lieu et d'en faire une reconstruction globale en temps réel.

La différence entre les systèmes exocentriques et égocentriques est grande, mais elle ne change pas de manière aussi apparente le type du système de télé-immersion comme la différence entre *outside-in* et *inside-out*. L'acquisition égocentrique offre une immersion efficace d'un visiteur sur le lieu, directement au sein de la zone sur laquelle son assistance est demandée. Les appareils d'acquisition embarqués dans les systèmes de téléprésence portables ou les robots de téléprésence permettent de capturer spécifiquement la région du lieu sur laquelle porte la collaboration. L'inconvénient est que la plupart de ces systèmes génèrent des scènes qui ne permettent de voir le lieu que d'un seul point de vue. Bien que la capture du lieu sous forme de scène 3D résolve ce problème, les méthodes de reconstruction 3D supposent souvent que le lieu est statique, étant donné la mobilité de la caméra. De plus, obtenir la représentation 3D de l'objet sur lequel la caméra est embarquée n'est pas trivial. En effet, l'absence ou le peu d'informations visuelles sur cet objet dans les données capturées complexifie alors l'obtention de sa reconstruction 3D. Ceci est préjudiciable par exemple pour développer un système de téléprésence portable qui capture une représentation de l'hôte. Actuellement, la télé-immersion 3D dans un lieu dynamique avec une acquisition égocentrique paraît alors plus difficile qu'avec une acquisition exocentrique. C'est pourquoi nous avons opté pour une caméra fixe plutôt que pour une caméra embarquée dans notre système de télé-immersion.

2.4 Inclusion de Scène

Après avoir capturé la scène distante, celle-ci doit être incluse dans le lieu de l'utilisateur. Cette inclusion est réalisée à l'aide de dispositifs d'immersion qui permettent d'immerger l'utilisateur dans la scène distante. Selon (Ohl, 2018), il existe deux types de dispositifs d'immersion : les dispositifs d'immersion matérielle et les dispositifs d'immersion virtuelle. Les dispositifs d'immersion matérielle sont utilisés dans les scenarios où la scène distante est représentée par un ou des artéfacts matériels sur le lieu de l'utilisateur. On parle d'extension matérielle ou de télé-immersion robotique. Un exemple peut être un bras robotique doté d'une ombre numérique avec une scène contenant alors les informations articulaires nécessaires pour animer le robot (Calandra *et al.*, 2022). Les dispositifs d'immersion virtuelle correspondent au cas où la scène distante est présentée grâce à des dispositifs de réalité étendue. Ces dispositifs peuvent être non-immersifs, semi-immersifs ou totalement immersifs (Boas, 2013). Des approches hybrides entre ces deux types de dispositifs peuvent aussi être envisagées. Cependant, on peut noter que la composition de la scène influence le type de dispositif d'immersion à utiliser. En effet, un dispositif d'immersion matérielle visant à présenter de manière matérielle un élément individuel de la scène, celui-ci va naturellement représenter un objet. Les interfaces à changement de formes, appelées *shape-changing display* ou *shape-changing interface* (Sturdee et Alexander, 2018; Rasmussen *et al.*, 2012), pourraient dans le futur ouvrir la voie à des extensions matérielles d'environnement en imaginant des technologies qui modifieraient physiquement la texture et la géométrie d'un lieu en fonction de la scène distante. Mais ces interfaces sont actuellement marginales pour les systèmes de télé-immersion, les dispositifs d'immersion matérielle ne sont pas adaptés pour immerger un utilisateur dans un environnement. De l'autre côté, un dispositif d'immersion virtuelle peut aussi bien être employé pour immerger un utilisateur aux côtés d'un objet ou au sein d'un environnement. Les dispositifs de réalité virtuelle faisant disparaître le lieu physique d'un utilisateur pour l'immerger dans un autre lieu, et ceux de réalité mixte intégrant un objet distant dans son lieu.

L'immersion d'un environnement permet de transporter l'utilisateur sur le lieu distant tandis que l'immersion d'un objet permet de ramener un objet distant dans son lieu. Ceci rappelle alors les différentes catégories de télé-immersion (figure 2.4). On note que la télé-immersion hôte-hôte et le télé-immersion visiteur-visiteur peuvent aussi bien se baser sur l'extension matérielle que sur l'extension virtuelle pour représenter les autres utilisateurs. En pratique, l'extension virtuelle est favorisée, surtout pour la télé-immersion visiteur-visiteur où le lieu commun est souvent virtuel. Quant à la télé-immersion hôte-visiteur, elle peut se baser sur une extension virtuelle pour l'hôte et le visiteur (réalité mixte et réalité virtuelle) ou alors l'extension matérielle avec l'extension virtuelle (le visiteur est représenté par un robot et immergé en réalité virtuelle).

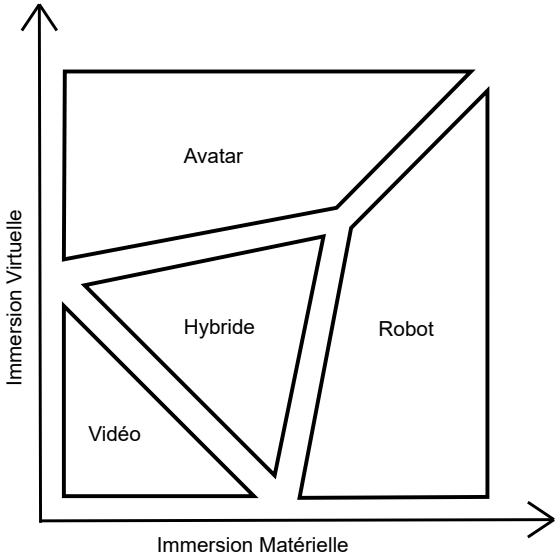


FIGURE 2.9 – Famille des types d'incarnation.

Comme expliqué, les utilisateurs d'une scène sont des éléments importants pour la télé-immersion et le type de télé-immersion désiré va impacter le dispositif à choisir pour les représenter. On parle d'incarnation pour décrire la technique d'immersion utilisée pour représenter une personne. Comme pour l'immersion, elle peut être matérielle à travers des robots anthropomorphiques ou virtuelle à travers des avatars. Une incarnation est qualifiée de télé-incarnation (Paulos et Canny, 2001, 1997) quand l'utilisateur est représenté par un robot anthropomorphique, et de virtuelle-incarnation quand l'utilisateur est représenté par un avatar. Mais cette simple classification en deux catégories ne suffit pas à représenter la diversité des incarnations de l'état de l'art. Nous proposons que les incarnations matérielles et virtuelles représentent plutôt deux continuums orthogonaux. Dans cet espace, nous avons identifié quatre familles d'incarnations avec des niveaux d'incarnations matérielles et virtuelles différents : les vidéos, les robots, les avatars et les hybrides (figure 2.9). Le choix de l'incarnation dépend du scénario de télé-immersion voulu (hôte-visiteur, hôte-hôte, visiteur-visiteur) mais aussi si l'interaction physique du visiteur est souhaitée (déplacer des éléments sur le lieu). Le choix de l'incarnation est aussi crucial pour la communication non-verbale car celle-ci dépend directement de l'incarnation des utilisateurs distants. Dans la suite, nous présentons des systèmes de télé-immersion utilisant ces types d'incarnations. Cette information sur l'incarnation est utilisée dans notre classification des systèmes de télé-immersion tableau A.1.

2.4.1 Incarnation Vidéo

1. `https://zoom.us`
2. `https://microsoft.com/microsoft-teams/teams-together-mode`

FIGURE 2.10 – Incarnations vidéos. Gauche : Zoom [1]. Milieu : Teams - Together Mode [2]. Droite : Visioconférence 2D dans une scène 3D (Hopkins et Benford, 1996).

Aujourd'hui, l'extension immersive la plus commune est la visioconférence. Cette technologie, largement utilisée au quotidien pour communiquer avec des personnes à distance, représente la forme de télé-immersion la plus simple. Dans la configuration classique, une webcam capture la scène sous forme de flux vidéo et l'élément d'intérêt est l'utilisateur distant. Les flux vidéos des différents utilisateurs sont présentés de manière non-immersive sur une interface en grille vidéo (figure 2.10 gauche). Cette interface peut être vue comme un lieu virtuel 2D abstrait sur lequel se déroule la télé-immersion. Même si ce type de logiciel est largement utilisé, leur utilisation répétée crée chez les utilisateurs un sentiment d'épuisement appelé *Zoom Fatigue* (Bailenson, 2021) qui peut dégrader la collaboration. De nouvelles interfaces ont alors été développées, toujours basé sur l'incarnation vidéo, afin de mitiger cet effet de fatigue. Une idée enrichissante pour la télé-immersion est de ne plus présenter les utilisateurs sous forme de grille vidéo mais de les rassembler sur un lieu commun virtuel. Le lieu commun peut alors être une image 2D (figure 2.10 milieu) ou une scène 3D dans laquelle les utilisateurs peuvent se déplacer (Dasari *et al.*, 2023; Young *et al.*, 2020; Hopkins et Benford, 1996) (figure 2.10 droite). Ce lieu commun est censé donner l'illusion aux utilisateurs qu'ils sont bien au même endroit. En particulier, il a été constaté que l'utilisation d'un lieu 3D commun permet un plus grand sentiment de présence sociale des utilisateurs (Hauber *et al.*, 2005). L'acquisition du lieu avec une caméra 360° peut aussi amener à des incarnations vidéo, comme avec l'appareil Microsoft RoundTable qui permet de tenir une visioconférence où les utilisateurs distants ont une vue panoramique sur tous les participants sur le lieu. L'incarnation d'un utilisateur est aussi une incarnation vidéo si la vidéo 360° est présentée en réalité virtuelle (Kim *et al.*, 2023; Jones *et al.*, 2020a; Rhee *et al.*, 2020). Nous classons aussi dans la catégorie d'incarnation vidéo des représentations plus abstraites de l'utilisateur. OmniGaze (Shiro *et al.*, 2018) est un système permettant de télé-immerger un visiteur sur un lieu avec plusieurs hôtes. Dans ce système, le lieu est capturé avec une caméra 360° permettant au visiteur de choisir librement la direction dans laquelle regarder. Sur le lieu commun, le visiteur n'est pas incarné par un flux vidéo, mais uniquement par la direction dans laquelle il est en train de regarder. Cette direction du regard est présentée grâce à une grille de LED recouvrant la caméra 360°, une LED

FIGURE 2.11 – Incarnations hybrides. Gauche MMSpace (Otsuka, 2016). Milieu : Beam [3]. Droite : Shader Lamps Avatar (Lincoln *et al.*, 2009).

allumée signifiant que le visiteur regarde dans cette direction. ThirdEye (Otsuki *et al.*, 2017) est un système de visioconférence pour les discussions entre deux utilisateurs qui utilise aussi l'idée de présenter la direction du regard avec un nouveau dispositif. En plus du flux vidéo de l'utilisateur distant, un artéfact matériel sous forme de globe oculaire est ajouté sur l'écran de l'utilisateur local. Ce globe oculaire est directement contrôlé par la direction du regard de l'utilisateur distant sur son écran, permettant alors de faire savoir à l'utilisateur local ce que l'autre est en train de regarder. Enfin, à la limite avec l'avatar 3D, GazeChat (He *et al.*, 2021) est une application de visioconférence sans flux vidéos mais avec des avatars créés à partir de photo des visages des utilisateurs. Sur l'interface, la direction du regard d'un avatar change en fonction de l'utilisateur qu'il est en train de regarder sur l'écran, servant ainsi à savoir quel utilisateur regarde quel autre pendant la conversation.

2.4.2 Incarnation Hybride

Une partie importante des propositions de l'état de l'art utilise des incarnations hybrides entre l'immersion matérielle et l'immersion virtuelle. Ces incarnations, ni avatars ni robots anthropomorphiques, sont principalement représentées par les robots de téléprésence. Généralement, un robot de téléprésence consiste en un écran pour incarner l'utilisateur sous forme de vidéo, combiné avec une base robotique téléopérée par l'utilisateur distant pour changer physiquement de point de vue. On parle de robot de téléprésence car les scènes créées permettent à un utilisateur qui en contrôle un de se sentir présent sur le lieu à la place du robot. Contrairement à la visioconférence, l'utilisateur d'un robot de téléprésence n'est pas immergé dans un lieu virtuel mais directement sur le lieu distant. Du point de vue de l'extraction de scène, cette différence s'explique par le fait que le dispositif d'acquisition (une caméra sur l'écran) ne sert plus uniquement à capturer la représentation d'un utilisateur mais l'environnement. La première

3. `https://telepresence.awabot.com`

catégorie de robot de téléprésence sont les *kinematic proxies* (Sirkin *et al.*, 2011) ou écrans mécaniquement mobiles (Kobayashi *et al.*, 2021). Il s'agit d'un écran monté sur une base mécanique fixe pouvant s'orienter dans différentes directions. Un utilisateur incarné dans ce robot peut alors contrôler dans quelle direction regarder en changeant l'orientation du bras mécanique. De nombreuses propositions de *kinematic proxies* ont été développées dans l'état de l'art (Adalgeirsson et Breazeal, 2010; Otsuka, 2016; Sakashita *et al.*, 2022) pour des tâches aussi variées que pour la tenue de réunions mixte ou la conception collaborative. La seconde catégorie de robot de téléprésence, plus répandue, sont les robots mobiles de téléprésence (*Mobile Robotic TelePresence*, *MRP*) (Kristoffersson *et al.*, 2013). Ces robots sont très similaires aux *kinematic proxies*, mais sont montés sur une base mobile permettant à l'utilisateur distant de se déplacer sur le lieu. Un robot de téléprésence mobile communément utilisé est Beam (figure 2.11 milieu), développé avec une base se déplaçant à l'aide de roues. Un robot de téléprésence peut aussi être équipé d'un pointeur ou d'un laser pour aider l'utilisateur distant à désigner un élément du lieu (Sakata *et al.*, 2003; Paulos et Canny, 2001). Mais il existe aussi des propositions plus originales, notamment avec des systèmes qui se déplacent non de manière terrestre mais aérienne à l'aide de dirigeables (Tobita *et al.*, 2011) ou de drones (Shakeri et Neustaedter, 2019). Le regard étant une information non-verbale essentielle, certaines incarnations hybrides se préoccupent de faire connaître précisément sa direction. En effet, la sensation que le portrait d'une personne nous regarde toujours dans les yeux, connue sous le nom d'effet Mona Lisa (Hecht *et al.*, 2014), fait qu'il est difficile de savoir dans quelle direction un utilisateur regarde à partir d'un flux vidéo sur un écran plat. Au lieu de se baser sur un simple écran comme la plupart des robots de téléprésence, Shader Lamps Avatar (Schubert *et al.*, 2012; Lincoln *et al.*, 2009) proposent une forme d'incarnation hybride intéressante, à la limite avec le robot anthropomorphique et l'avatar. Ce système utilise un visage humanoïde animatronique à l'échelle, sur lequel est affiché le visage de l'utilisateur distant grâce à des projecteurs (figure 2.11 droite), permettant ainsi de créer une représentation 3D réaliste de celui-ci. LiveMask (Misawa *et al.*, 2012a) reprend cette idée dans un système avec lequel la vidéo de l'utilisateur distant est affiché en sur un écran mobile en forme de visage. Cet ajout de relief sur la vidéo permet de résoudre le problème de l'effet Mona Lisa.

2.4.3 Incarnation Robotique

La téléopération consiste à contrôler une machine à distance. On parle d'incarnation robotique lorsqu'un utilisateur contrôle à distance un robot anthropomorphique, avec notamment un visage et des bras. Cette incarnation robotique inclut alors toutes les formes de téléopération de robots humanoïdes (Darvish *et al.*, 2023). L'incarnation robotique s'appuie largement sur l'utilisation d'artéfacts matériels pour représenter un

FIGURE 2.12 – Incarnations robotiques. Gauche : TELESAR VI (Tachi *et al.*, 2020). Milieu : Geminoid H (Nishio *et al.*, 2007). Droite : EDGAR (Ching *et al.*, 2016).

visiteur. Le contrôle des artéfacts matériels a l'avantage de permettre aux utilisateurs distants de réaliser des tâches physiques sur le lieu commun comme déplacer des objets. Aujourd'hui, un des plus complets est le robot TELESAR VI (Tachi *et al.*, 2020) (figure 2.12 gauche) avec 67 degrés de liberté. Pour avoir une communication la plus efficace entre les utilisateurs distants et sur site, et notamment transmettre fidèlement les comportements non-verbaux, des robots réalistes d'humains ont été mis au point. Les plus connus sont les robots Geminoid H (figure 2.12 milieu) et Geminoid F (Nishio *et al.*, 2007), produits à partir de modèles humains. Ces robots sont capables de reproduire des expressions faciales humaines grâce à une animation des yeux et de la bouche. Le problème de ce type robot est qu'il ne peut représenter qu'une unique identité, les différents utilisateurs de ce système étant tous incarnés sous la même apparence. Pour résoudre ce problème, les robots TELESAR II (Tachi *et al.*, 2008) et EDGAR (Ching *et al.*, 2016) (figure 2.12 droite) proposent d'afficher directement sur le robot un flux vidéo de l'utilisateur distant, permettant ainsi de changer l'apparence du robot en fonction de l'utilisateur le contrôlant. En plus de pouvoir physiquement manipuler des éléments du site, cette incarnation offre aussi des avantages pour la collaboration entre les utilisateurs. Un robot anthropomorphique réaliste permet entre autres de susciter un plus grand sentiment de présence qu'une représentation vidéo (Sakamoto *et al.*, 2007). Une incarnation robotique procure aussi plus de confiance qu'une incarnation sous forme d'avatar 3D (Pan et Steed, 2016). Cependant, l'incarnation en robot anthropomorphique est la plus sensible au phénomène de vallée de l'étrange, ou *uncanny valley*, qui rend les imperfections de l'apparence du robot flagrantes plus il ressemble à un humain (Mori *et al.*, 2012). Il est alors possible que ce sentiment d'étrangeté de l'incarnation résulte en une communication moins naturelle. Surtout, l'inconvénient majeur de cette incarnation est qu'elle nécessite un système physique complexe difficilement transportable et

FIGURE 2.13 – Incarnations avatars. Gauche : Starline (Lawrence *et al.*, 2021). Milieu : Holoportation (Orts-Escolano *et al.*, 2016). Droite : VROOM (Jones *et al.*, 2020a).

donc pas idéal pour un système de télé-immersion mobile.

L'état de l'art comporte aussi des systèmes difficiles à séparer entre les incarnations hybrides et robotiques. Par exemple, (Onishi *et al.*, 2016) ont mis au point un système combinant la vidéo avec un bras robotique anthropomorphique contrôlé par l'utilisateur distant pour renforcer le sentiment de présence. Nous incluons aussi dans l'incarnation robotique les quelques systèmes recourant à des interfaces à changement de formes. Ces systèmes représentent un utilisateur grâce à un support matériel, qui n'est pas un robot anthropomorphique, accompagné de la vidéo. Nous considérons cette incarnation comme robotique pour souligner son aspect physique et matériel. PopObject (Kushida et Nakanishi, 2018) est un système de télé-immersion vidéo avec lequel la représentation de l'utilisateur distant est diffusée sur un écran déformable. Le système est doté d'un écran avec un mécanisme d'extension qui permet de le pousser de pour rajouter du relief à la vidéo. L'écran n'est donc pas plat mais plus ou moins avancé en fonction de la proximité de l'élément avec la caméra qui le filme. L'incarnation d'un utilisateur sur cet écran en relief semble améliorer le sentiment de présence. (Leithinger *et al.*, 2014) proposent un système où l'utilisateur distant est incarné par une représentation vidéo avec une interface à changement de formes pour modéliser de manière physique ses mains. Cette interface à changement de formes consiste en une grille de pièces en plastique, chacune pouvant être actionnée individuellement afin d'être montée ou descendue dans le but d'avoir une surface en relief. Pour ce faire, les mains de l'utilisateur distant sont posées sur une table et une caméra de profondeur au plafond va transmettre à l'utilisateur sur site le relief à produire sur l'interface à changement de formes. L'intérêt est que l'utilisateur distant interagit physiquement avec des objets sur la table à distance, comme déplacer une balle.

2.4.4 Incarnation Avatar

Le dernier type d'incarnation est celle sous forme d'avatar. La différence avec l'incarnation vidéo est que la représentation sous forme d'avatar est une représentation 3D avec des informations géométriques, contrairement à la vidéo qui est une représentation

2D. Cette incarnation repose largement sur kes dispositifs de réalité étendue pour afficher ces informations 3D. Une configuration courante dans la littérature est d'incarner les utilisateurs distants à l'aide d'écrans. Certains systèmes ont la particularité de se baser sur des écrans à effet holographique permettant de voir objets en 3D avec une parallaxe de mouvement, c'est-à-dire de pouvoir voir l'objet sous un nouveau point de vue en changeant physiquement de position. Cette technologie est à la base de Starline (Lawrence *et al.*, 2021), un projet visant à faire communiquer deux utilisateurs distants à travers un écran comme s'ils étaient dans la même pièce en tête-à-tête (figure 2.13 gauche). Le système consiste en une cabine avec un écran pour donner l'illusion que l'utilisateur distant est dans la même pièce derrière une vitre. (Wang *et al.*, 2023b) proposent un système similaire mais sans espace dédié. Ces systèmes se basent sur un ensemble de caméras autour de l'écran, permettant de reconstruire un avatar de l'utilisateur et de suivre la position du regard pour déterminer le point de vue à afficher pour le rendu de l'autre utilisateur sur l'écran. D'autres propositions affichent l'avatar distant à travers un écran pour donner l'illusion que l'utilisateur distant est présent sur le lieu en utilisant un écran semi-transparent ou un écran reproduisant l'arrière-plan pour donner une apparence transparente (Alvarez *et al.*, 2022; Plüss *et al.*, 2016). D'autres utilisent des écrans sur mesure comme TeleHuman2 (Gotsch *et al.*, 2018) qui se base sur un écran holographique cylindrique donnant la possibilité de voir l'utilisateur distant tout autour de l'écran. HoloKinect (Siemonsma et Bell, 2022) est un système qui incarne aussi un utilisateur avec un écran, mais avec la particularité d'afficher l'utilisateur distant sur un écran holographique portatif. D'autres formes d'écrans holographiques plus atypiques peuvent aussi afficher l'avatar d'un utilisateur en 3D (Córdova-Esparza *et al.*, 2019; Jones *et al.*, 2009). Certains systèmes incarnent des utilisateurs sur un lieu virtuel grâce à un CAVE (Roberts *et al.*, 2015; Beck *et al.*, 2013; Steed *et al.*, 2012; Gross *et al.*, 2003). Par exemple, VirtualCube (Zhang *et al.*, 2022b) est un système de télé-immersion incarnant chaque utilisateur dans un lieu commun dans le but de simuler une discussion en face à face. Les utilisateurs distants sont ramenés dans le lieu grâce à un mur d'écrans entourant la pièce, et sont virtuellement installés à une table avec l'utilisateur local. Les approches de réalité augmentée avec un casque sont aussi utilisées pour incarner l'avatar sur le lieu commun. Le plus célèbre des systèmes, Holoportation (Orts-Escolano *et al.*, 2016), ajoute dans le lieu un avatar 3D photoréaliste de l'utilisateur distant à l'aide d'un Hololens (figure 2.13 milieu). Dans AVT (Rhee *et al.*, 2020), le visiteur incarne un avatar simpliste avec une scène capturée par une caméra 360°. Le visiteur est immergé en réalité virtuelle dans la vidéo 360° tandis que l'hôte voit l'avatar du visiteur en surimpression à la position de la caméra. Mais les techniques de réalité augmentée par projection (*projected augmented reality*) (Mamone *et al.*, 2021) sont aussi utilisées pour incarner un utilisateur sur le lieu. Room2Room (Pejsa *et al.*,

2016) est une proposition amenant un utilisateur distant sur un lieu local avec cette approche. Une caméra RGB-D capture la représentation 3D de l'utilisateur et un projecteur affiche cette représentation en surimpression sur l'autre lieu, se passant ainsi de l'utilisation d'écrans.

Les dispositifs d'immersion virtuelle peuvent être combinés avec les supports matériels des approches hybrides (sous-section 2.4.2) pour permettre aux utilisateurs distants de se déplacer sur le lieu. VROOM (Jones *et al.*, 2020a) est un système où le visiteur est incarné par un robot de téléprésence mobile qui est contrôlé en réalité virtuelle. L'hôte, grâce à un casque de réalité mixte, voit l'avatar du visiteur en surimpression sur le robot de téléprésence (figure 2.13 droite). Holobot (Kim *et al.*, 2023) propose un système similaire à la différence que l'avatar du visiteur est visualisé grâce à un écran holographique directement monté sur le robot. Enfin, Viewpoint-Controllable Telepresence (Luo *et al.*, 2023) est un système dans lequel le visiteur est incarné, en réalité virtuelle, à l'aide d'une caméra stéréoscopique monté sur un bras robotique servant de *kinematic proxy*. Ce système permet à un hôte de voir le visiteur en surimpression sur le robot avec un casque de réalité mixte.

L'incarnation sous forme d'avatar 3D peut être plus avantageuse qu'une incarnation vidéo pour la collaboration entre des hôtes et des visiteurs. (Campbell *et al.*, 2020) ont réalisé une comparaison entre les réunions en visioconférence et les réunions en réalité virtuelle (avec une incarnation en avatar) et ont montré que les réunions en réalité virtuelle suscitent un plus grand sentiment de présence et de proximité physique. Cette incarnation peut aussi être celle qui permet d'avoir la communication non-verbale la plus proche que celle observée lors de discussions en face à face (Smith et Neff, 2018). Ce type d'incarnation est le plus intéressant dans le développement d'un système de télé-immersion mobile, car il permet de représenter des utilisateurs en 3D à l'aide de dispositifs d'immersion portables comme des casques de réalité mixte.

2.5 Conclusion

Dans ce chapitre, nous avons défini la télé-immersion dans un nouveau cadre théorique soulignant les différences fondamentales qui existent entre les systèmes. Dans notre cadre, chaque utilisateur d'un lieu est un hôte ou un visiteur. Un hôte amène auprès de lui dans son lieu un autre utilisateur, tandis qu'un visiteur est virtuellement téléporté sur un autre lieu. Cette distinction met en évidence le lieu où se déroule la collaboration, appelé lieu commun. La différence entre hôte et visiteur permet aussi d'identifier trois types de télé-immersion : hôte-visiteur où il y a au moins un hôte et un visiteur, hôte-hôte où il n'y a que des hôtes et visiteur-visiteur où il n'y a que des visiteurs. Le concept de symétrie apparaît alors naturellement. Un système est symétrique s'il n'y a que des hôtes ou que des visiteurs, sinon il est asymétrique. Une autre notion importante est

la scène qui correspond à la représentation des éléments d'intérêt d'un lieu. À travers une scène, un utilisateur accueille un utilisateur d'un autre lieu, ou éprouve la sensation d'être sur un autre lieu. La télé-immersion repose alors sur une étape d'extraction de scène pour capturer les éléments d'intérêt du lieu, et sur une étape d'inclusion de scène pour présenter les éléments d'intérêt du lieu à l'utilisateur. Notre cadre théorique insiste sur le fait qu'une scène ne contient que deux types d'éléments : objet et environnement. Un objet est une représentation précise d'un élément du lieu comme le scan 3D d'une personne, tandis que l'environnement est une représentation globale du lieu comme une image 360°. Une propriété essentielle que nous en tirons est que le lieu commun dépend uniquement de l'environnement de la scène. Enfin, les personnes d'un lieu étant un type d'objet particulier, on appelle incarnation spécifiquement leurs représentations dans la scène. Notre théorie propose quatre familles d'incarnation : vidéo, hybride, robot et avatar. Celles-ci correspondent respectivement aux logiciels de visioconférence, aux robots de téléprésence, aux robots humanoïdes et aux dispositifs de réalité étendue.

Les notions introduites dans notre cadre théorique permettent de formaliser le système de télé-immersion 3D vers lequel nous souhaitons tendre (section 1.3). Notre système est asymétrique, avec des hôtes qui accueillent des visiteurs totalement immergés sur un lieu dynamique. Dans notre cas d'usage d'enseignement à distance, le lieu commun est la salle de classe, le professeur et étudiants sur site sont les hôtes, et les étudiants à distance sont les visiteurs. Pour l'acquisition, nous exploitons une caméra 360°, un dispositif *inside-out* facilement transportable. Le choix d'un dispositif *inside-out* est judicieux car ils sont la base de la plupart des systèmes asymétriques. Néanmoins, pour que le système donne aux visiteurs une liberté de navigation, une reconstruction 3D du lieu doit être obtenue. Or, les dispositifs *inside-out* ne sont pas adaptés à cette tâche. Cette reconstruction est d'autant plus dur à obtenir avec une caméra statique. Quant à l'immersion, notre système repose sur des incarnations en avatars avec des dispositifs de réalité mixte. Ce choix est justifié par la dimension nomade de notre système où la télé-immersion pourra prendre place sur un nouveau lieu simplement en déplaçant la caméra 360° avec un dispositif d'immersion léger (casque de réalité mixte, pas d'immersion robotique ou hybride). Le prochain chapitre initie développement de notre système de télé-immersion en se basant exclusivement sur les données d'une caméra 360°.

Chapitre 3

Télé-Immersion 360°

Pour développer notre système de télé-immersion, nous avons choisi la caméra 360° comme brique de base. Notre ambition est de pouvoir simplement poser une caméra 360° dans un lieu d'intérêt et que des utilisateurs sur place puissent se réunir sur ce lieu avec des utilisateurs distants. Un tel système est désirable pour une plateforme d'enseignement à distance : le professeur place la caméra 360° dans la salle de classe et les étudiants suivent le cours de chez eux comme s'ils étaient physiquement présents. L'étude de ce dispositif étant primordiale pour vérifier si ce système est réalisable, nous étudions dans ce chapitre ce que sont les caméras et images 360°. À partir de ces informations, nous proposons un premier système de télé-immersion relativement simple pour immerger des utilisateurs dans un lieu capturé par une caméra 360° statique. Cette première proposition vise à juger les possibilités de la caméra 360° sans capture 3D pour la télé-immersion, et sa pertinence pour réaliser des cours en ligne immersifs.

3.1 Image 360°

En 1787, le peintre Robert Barker dépose un brevet pour son invention *la nature à coup d'œil* (Belisle, 2015). Son invention consiste en de grandes peintures de paysages circulaires sans frontières qui pouvaient être observées à partir d'un point central (Kamcke et Hutterer, 2015; Lescop, 2017). Renommé par la suite *panorama*, du grec *pan* pour tout et *horama* pour vue, ses peintures cherchaient à donner l'illusion au spectateur d'être à l'intérieur de la scène. Quelques années plus tard, son panorama de Londres, exposé sur la surface intérieure d'une pièce cylindrique et visible depuis une plate-forme au milieu de la pièce, fut un grand succès populaire. Cette idée d'image circulaire sans frontières fait des peintures de Barker l'ancêtre de l'image 360°. Il faut attendre 1843 pour que la peinture soit remplacée par un dispositif capturant automatiquement un panorama avec la première caméra panoramique de Puchberger d'Autriche (Benosman et Kang, 2001). Cependant, son modèle de caméra était statique, ce qui

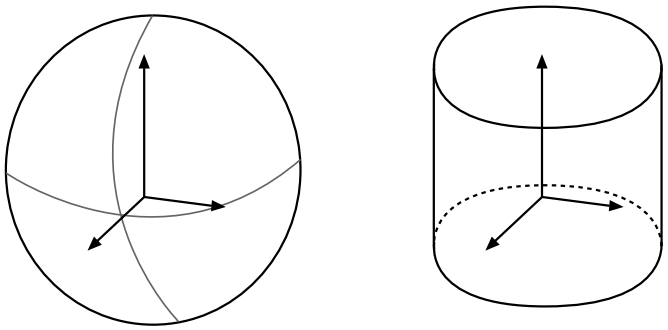


FIGURE 3.1 – Différence entre image omnidirectionnelle et image panoramique, tiré de (Corbillon, 2019). Gauche : Image omnidirectionnelle, sphérique. Droite : Image panoramique, cylindrique.

limitait son champ de vision à 150°. Garella d'Angleterre propose en 1857 un modèle amélioré avec une caméra rotative capable de capturer un panorama complet à 360°. La caméra 360° moderne est finalement inventée en 1970 (Gledhill *et al.*, 2003) avec pour avantage d'être plus rapide que les modèles précédents, permettant ainsi de capturer des vidéos 360°. Aujourd'hui, la caméra 360° s'est largement généralisée et sa capacité de capturer des images et vidéos dans lesquels une personne peut être immergée fait de cet appareil une base judicieuse pour notre système de télé-immersion. Dans cette section, nous présentons les propriétés des caméras 360° et des données qu'elles capturent.

3.1.1 Image Omnidirectionnelle

Une image 360°, ou une vidéo 360° par extension, désigne une projection de la surface d'une sphère vue depuis son centre (Lee *et al.*, 2016). Le contenu d'une image 360° couvre l'ensemble du champ de vision à (360 x 180)° contrairement à une image perspective traditionnelle qui ne couvre qu'un plan limité (Xu *et al.*, 2020). Les images 360° sont aussi appelées images sphériques, images panoramiques ou images omnidirectionnelles. Cependant, les images 360° englobent aussi les images cylindriques qui ne contiennent pas d'informations près des pôles de la sphère et donc ne couvrent pas verticalement l'ensemble du champ de vision à 180°. La figure 3.1 illustre cette différence. Pour éviter la confusion, nous utilisons le terme omnidirectionnel pour décrire les images et vidéos 360°, et réservons le terme panoramique pour les images et vidéos cylindriques (Corbillon, 2019). Une caméra est aussi dite omnidirectionnelle si elle capture des images et vidéos omnidirectionnelles. Lors du visionnage d'une image omnidirectionnelle, le spectateur contrôle la direction dans laquelle il regarde la scène à travers la fenêtre de visionnage, c'est-à-dire la projection d'une portion de la sphère sur un plan (Shafi *et al.*, 2020). Cette fenêtre de visionnage est contrôlée à travers 3 degrés de liberté (*3-DoF*) : l'angle de roulis, l'angle de tangage et l'angle de lacet.

Caméra Omnidirectionnelle

Comme discuté au chapitre précédent, une caméra omnidirectionnelle appartient à la famille des dispositifs *inside-out*. Mais les caméras omnidirectionnelles peuvent être divisées en deux catégories : les dispositifs mono-point de vue et les dispositifs multi-point de vue. Plus formellement, les premiers sont les modèles à point de vue unique (*Singular Viewpoint*, *SVP*) ou monocentriques, tandis que les seconds sont les modèles à point de vue non-unique (*non-Singular Viewpoint*, *non-SVP*) ou polycentriques (Gurrieri et Dubois, 2013). Lors de la capture d'un lieu avec un modèle SVP, les rayons lumineux convergent vers un point unique du dispositif. Les caméras SVP sont alors des dispositifs mono-caméra, englobant les caméras dioptriques qui utilisent des lentilles déformantes (lentille *fisheye*) et les caméras catadioptriques qui utilisent un miroir déformant (miroir parabolique, hyperbolique, elliptique) (Scaramuzza, 2014). Ces caméras ont l'avantage d'être peu coûteuse car elles combinent simplement une caméra classique avec des lentilles ou des miroirs, et la calibration est simplifiée car seule la déformation des lentilles ou des miroirs doit être modélisée (Fan *et al.*, 2019). Ces caméras ont tout de même l'inconvénient d'avoir une résolution limitée et de générer du flou sur certaines régions de l'image à cause d'un échantillonnage non-uniforme des rayons lumineux. Une caméra à rotation autour de son point de convergence est aussi une caméra SVP (Konrad *et al.*, 2017). Dans le modèle non-SVP, les rayons lumineux ne convergent pas vers un unique point mais vers plusieurs. Le lieu n'est pas strictement capturé depuis un unique point de vue mais depuis plusieurs points de vue proches. Ces différentes images des différents points de vue sont assemblées en une image grâce à une étape de *stitching* pour simuler l'observation depuis un unique point de vue. Les caméras non-SVP sont généralement des dispositifs multi-caméras, ou caméras polydioptriques, qui utilisent plusieurs caméras avec des champs de vision se chevauchant. Les caméras non-SVP incluent aussi les dispositifs à unique caméra rotative, avec un axe de rotation décentrée de son point de convergence. Ces caméras ont l'avantage de produire des images avec une résolution et une qualité supérieure mais elles peuvent présenter des artéfacts de discontinuités aux jonctions des sous-images lorsque le *stitching* n'est pas optimal.

Projection

Pour transmettre et visionner des images ou vidéos omnidirectionnelles, il est nécessaire de transformer la surface sphérique en un plan. Une projection est la transformation de la sphère vers un plan. Les projections sont nombreuses, chacune ayant ses avantages, mais celles-ci sont classées dans la littérature en deux catégories : les projections à qualité uniforme (indépendante de la fenêtre de visionnage) et les projections à qualité variable (dépendante de la fenêtre de visionnage) (Chen *et al.*, 2018). Les projections à qualité uniforme convertissent une image omnidirectionnelle avec une

qualité homogène sur l'ensemble de l'image (indépendamment de la région visualisée par le spectateur). La plus courante des projections est la projection équirectangulaire (*equirectangular projection*, *ERP*). La projection équirectangulaire est une fonction linéaire faisant correspondre directement les coordonnées horizontales i et verticales j d'un pixel respectivement à la longitude φ et à la latitude θ de la sphère avec la formule suivante :

$$\begin{aligned} i &= w\,(\frac{1}{2} + \frac{\varphi}{2\pi}) \\ j &= h\,(\frac{1}{2} - \frac{\theta}{\pi}) \end{aligned} \tag{3.1}$$

où w et h sont la largeur et la hauteur de l'image, avec l'origine de l'image en haut à gauche. Étant donné que $\varphi \in [-\pi, \pi]$ et $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, une convention pour les images équirectangulaires est d'avoir w valant le double de h. Cette projection est la représentation standard d'une image omnidirectionnelle grâce à sa simplicité. Néanmoins, elle a pour inconvénient de suréchantillonner la sphère aux pôles, occasionnant des pixels redondants et des distorsions. La variante *EAP* (*equal area projection*) a été proposée pour réduire ce suréchantillonnage à l'aide d'un coefficient dépendant de la latitude, mais elle accentue les distorsions. Une autre projection standard est la projection *cubemap* (*CMP*). L'idée est de projeter la sphère sur un cube circonscrit et d'assembler les six faces du cube sur une même image. Les faces du cube peuvent être assemblées sur la même image en suivant différents agencements, résultant en des taux de compression plus ou moins élevés. Cette projection ne présente pas les distorsions de la projection équirectangulaire, mais crée des discontinuités entre les faces du cube et suréchantillonne près des coins. La projection *cubemap* se généralise à des solides circonscrits quelconques avec les projections en patchs (*patch-based projection*). Ces solides peuvent être des dodécaèdres (*dodecahedron projection*), des octaèdres (*octahedron projection*, *OHP*) ou des icosaèdres (*icosahedron projection*, *ISP*). Le solide est choisi de manière à équilibrer le nombre de face : plus de faces implique une réduction du suréchantillonnage mais aussi une augmentation des discontinuités.

De l'autre côté, les projections à qualité variable permettent de convertir l'image omnidirectionnelle avec une qualité dépendante de la région visualisée. L'idée est de diminuer la qualité de l'image omnidirectionnelle aux régions où le spectateur ne regarde pas. Ces projections sont particulièrement utiles pour la transmission de vidéo omnidirectionnelle afin de réduire la taille des données pour une même qualité visuelle dans la fenêtre de visionnage. Une projection de ce type est la projection pyramidale. La projection pyramidale est une projection en patchs qui centre la projection sur la fenêtre de visionnage. La sphère est projetée sur une pyramide circonscrite, la fenêtre de visionnage est projetée sur la base de la pyramide, la seule face échantillonnée en pleine

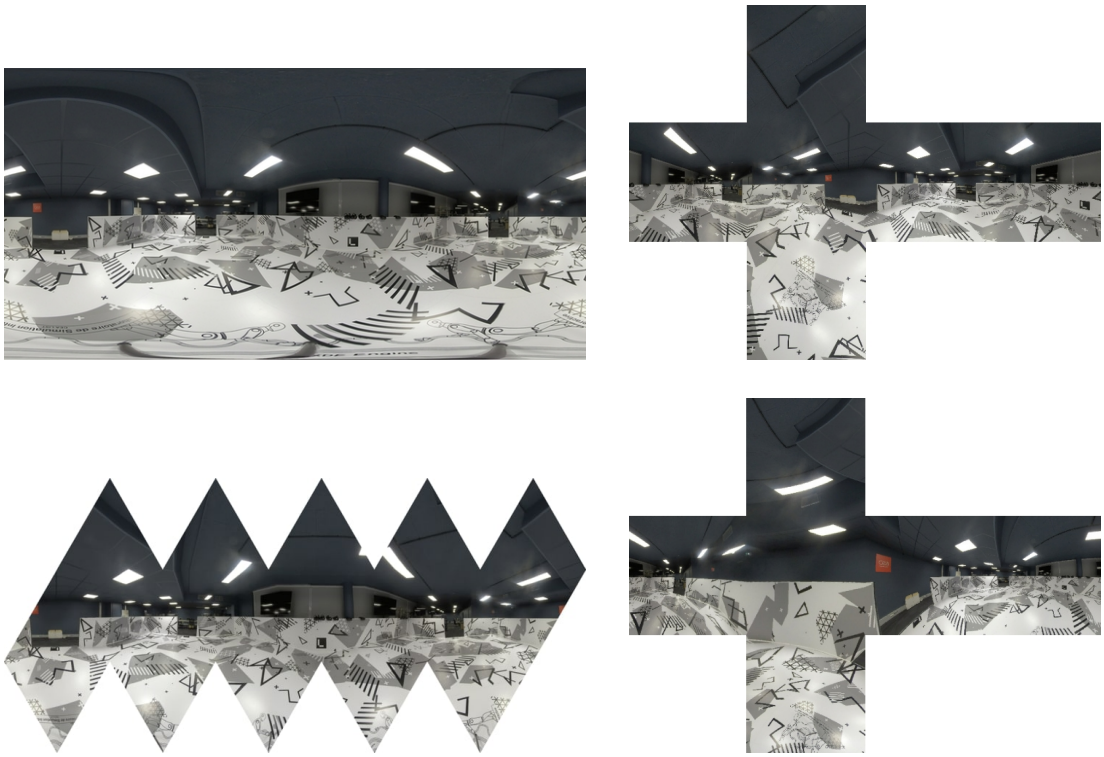


FIGURE 3.2 – Projections d'une image omnidirectionnelle. Haut-Gauche : Projection équirectangulaire. Haut-Droite : Projection *cubemap*. Bas-Gauche : Projection en patchs sur un icosaèdre (*ISP*). Bas-Droite : Projection *offset-cubemap*.

résolution, et le reste de la sphère est projeté sur les autres faces de la pyramide. Une variante est de projeter la sphère sur une pyramide carrée tronquée (un tronc à base carré). Enfin, une dernière projection à qualité variable est la projection *offset-cubemap* (Zhou *et al.*, 2017), une variante de *cubemap*. Dans cette projection, le centre de la sphère et du cube circonscrit ne sont pas les mêmes, le centre de la sphère est décalé de celui du cube pour que les faces du cube couvrent des portions de tailles différentes de la sphère. Avec cette projection, la dégradation de la qualité paraît plus progressive pour le spectateur qu'avec une projection pyramidale.

Une dernière projection pertinente à citer est la projection gnomonique (Monteleone *et al.*, 2018). La projection gnomonique permet de projeter seulement une portion de la sphère vers un plan tangent à la sphère sur cette portion. Cette projection est utilisée pour obtenir la fenêtre de visionnage quand le support n'affiche pas totalité de l'image omnidirectionnelle en même temps, par exemple avec un casque de réalité virtuelle. Malgré toute la diversité des projections, aucune n'est strictement supérieure aux autres. Chaque projection entraîne inévitablement des distorsions qui ne peuvent pas être corrigées. Ce résultat est la conséquence du *theorema egregium* de Gauss qui

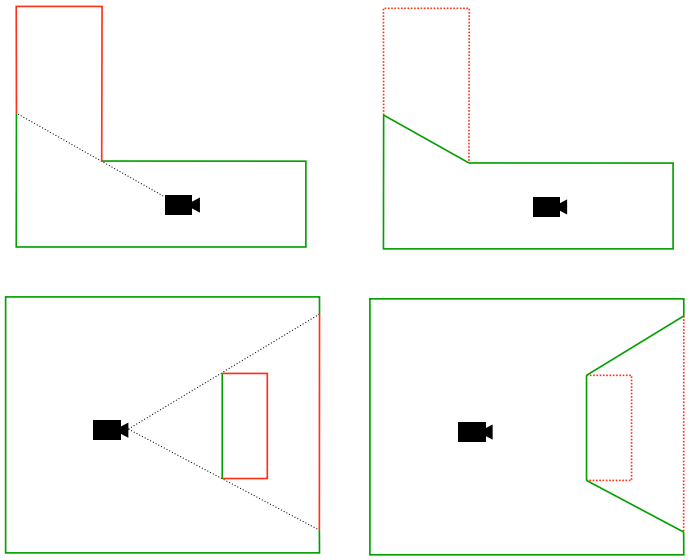


FIGURE 3.3 – Capture d'un lieu avec une caméra omnidirectionnelle. En vert les régions visibles par la caméra, en rouge les régions occultées. Haut : Le lieu n'est pas entièrement visible par la caméra omnidirectionnelle (gauche), les éléments visibles devant tous être connectés à leurs voisins. Une nouvelle surface est donc créée ayant pour conséquence de boucher un trou (droite). Bas : Le lieu contient une composante connexe indépendante introduisant des occultations (gauche), la caméra ajoute de nouvelles surfaces entre éléments indépendants. Le nombre de composantes connexe est réduit à 1 (droite).

implique qu'une surface sphérique n'est pas isométrique à un plan (Eder *et al.*, 2020).

3.1.2 Image Omnidirectionnelle et Topologie

Une caméra omnidirectionnelle étant un dispositif d'acquisition qui permet de capturer les informations de lumière à une position donnée dans toutes les directions. Les propriétés de la lumière font qu'un point est visible par la caméra s'il existe un chemin entre ce point et la position de la caméra, le chemin étant un segment en condition normale. L'ensemble des points visibles peuvent composer une ou plusieurs surfaces sur le lieu réel, mais la caméra ne pouvant pas capturer d'informations topologiques (comme les trous ou les composantes connexes) la topologie entre les points visibles est inconnue. Cette ambiguïté fait que l'ensemble des points visibles sont considérés comme appartenant à une même unique surface où chacun des points est connecté à ses voisins. Ainsi, des trous peuvent se retrouver bouchés ou des éléments distincts agrégés dans la même composante connexe comme illustré figure 3.3. Ces nouvelles surfaces bouchant les trous et agrégeant les composantes connexes, qui n'existent pas en réalité, sont appelées surfaces fantômes (Kanade *et al.*, 1997). Ces surfaces sont directement la conséquence de l'hypothèse que les points sont tous connectés. Ces surfaces fantômes ne sont pas visibles tant que la scène est visualisée à la position de la caméra comme avec

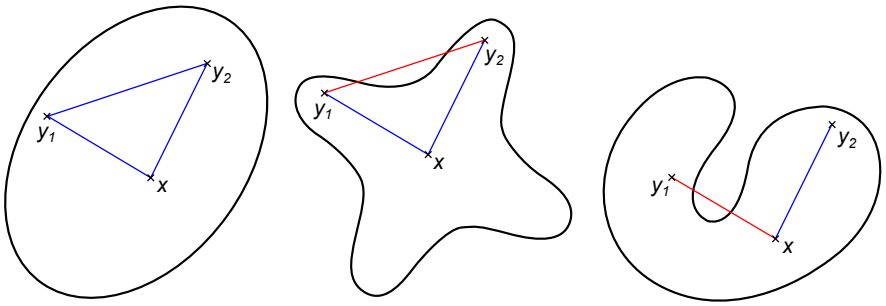


FIGURE 3.4 – Convexe, convexe étoilé et non convexe étoilé en 2D, tiré de (Minja et Šenk, 2019). Gauche : Convexe, tous les points sont des centroïdes. Pour toute paire de point, le segment est à l'intérieur de D. Milieu : Convexe étoilé, il y a au moins un centroïde et il n'est pas forcément unique. Le segment entre un point quelconque et un centroïde est à l'intérieur de D. Droite : Pas un convexe étoilé, il n'y a pas de centroïde.

un visionneur d'image omnidirectionnelle classique. Mais l'introduction de la 3D génère à ces régions des artéfacts qui peuvent dégrader le confort du spectateur (Dupont de Dinechin et Paljic, 2018). Cette hypothèse de connexité entre tous les voisins permet surtout de considérer qu'une caméra omnidirectionnelle capture alors l'enveloppe d'un convexe étoilé.

Un ensemble $D \subseteq \mathbb{R}^n$ est un domaine étoilé ou convexe étoilé s'il existe un point $x \in D$ tel que le segment entre x et un point quelconque $y \in D$ appartient aussi à D (Minja et Šenk, 2019; Lien, 2007). Formellement, le segment $[x, y] \subseteq D \Leftrightarrow (1-v)x + vy \in D$ pour $0 \leq v \leq 1$ (Kawana *et al.*, 2020). On dit que D est convexe par rapport à x, ou que x est le centroïde de D (Zhang et Lu, 2005). Un exemple de comparaison entre un convexe, un convexe étoilé et une figure quelconque en 2D est donné figure 3.4. Dans une image omnidirectionnelle, la caméra joue le rôle centroïde et capture la surface d'un convexe étoilé. Si on a un point tel que le chemin entre ce point et le centroïde n'est pas inclus dans le convexe étoilé, alors le point n'est pas visible dans le cas. C'est le cas par exemple si on a un mur entre la caméra et le point. En particulier, un convexe étoilé ne permet de modéliser qu'une unique composante connexe, ce qui explique pourquoi des éléments indépendants ne peuvent pas être distingués. Précisément, la caméra omnidirectionnelle permet de capturer un convexe étoilé particulier : l'enveloppe étoile-convexe intérieure du lieu. Soit un lieu $S \subseteq \mathbb{R}^n$, $D_p \subseteq \mathbb{R}^n$ est l'enveloppe étoile-convexe intérieure de S en p si D_p est le plus grand convexe étoilé de centroïde p contenu dans S. Avec p correspondant à la position de la caméra, D_p correspond à ce que capture la caméra en p. La surface visible figure 3.3 (droite) correspond exactement à l'enveloppe étoile-convexe intérieure centrée sur la position de la caméra. La caméra omnidirectionnelle capture l'enveloppe étoile-convexe intérieure d'un lieu car ce sont les plus longs segments possibles qui sont capturés par la caméra, la lumière rebondissant sur les enveloppes des éléments de la scène. En général, c'est toujours le plus long segment qui est capturé s'il

n'y a pas de changement de milieu de propagation de la lumière (comme de la fumée).

Cette définition de ce que capture une caméra omnidirectionnelle permet de dégager une propriété intéressante. Si le lieu est un convexe étoilé de centroïde p et que la caméra omnidirectionnelle est posée en p alors l'ensemble du lieu est visible. Ceci est la conséquence du fait que l'enveloppe étoile-convexe intérieure du lieu en p est le lieu entier si le lieu est un convexe étoilé de centroïde p. Si la caméra n'est pas posée sur un centroïde du lieu, alors l'enveloppe étoile-convexe intérieure centrée sur la position de la caméra ne couvre pas l'ensemble du lieu. Mais si la caméra est sur un centroïde, alors une unique caméra est suffisante pour capturer l'ensemble du lieu sans occultations. Cette situation est exploitée au chapitre 4 pour avoir un système avec une navigation libre sur un lieu simple. Mais en pratique, cette contrainte est rarement respectée : soit le lieu n'a pas une topologie en convexe étoilé (figure 3.3 haut) ou le lieu est composé de plusieurs composantes connexes (figure 3.3 bas). Entre ces deux cas de figures, le deuxième parait plus facile à résoudre. Le chapitre 5 propose une solution pour ce cas. Une autre propriété est qu'un convexe étoilé est paramétrisé par une fonction appelée fonction de rayon (Granström *et al.*, 2017). Cette fonction fait correspondre à un ensemble d'angles le rayon entre le centroïde et le point sur la surface pour l'ensemble d'angles donné. En 2D, un convexe étoilé est paramétrisé par une fonction d'angle à un seul paramètre, et en 3D par une fonction d'angle à deux paramètres. La carte de profondeur omnidirectionnelle, introduite au chapitre 4, est une discrétisation de cette fonction de rayon. Pour conclure du côté de la télé-immersion, il est important de noter que cette topologie capturée par une caméra omnidirectionnelle rend l'image omnidirectionnelle particulièrement adaptée pour représenter un environnement. Un environnement consistant en des éléments à un champ lointain, la possibilité pour un utilisateur de se déplacer autour d'un élément de l'environnement est limitée. Cette contrainte est imposée par le fait qu'un convexe étoilé ne peut représenter qu'une unique composante connexe. Une image omnidirectionnelle peut toutefois être utilisée pour modéliser un objet 3D, c'est-à-dire un élément vu de l'extérieur, contrairement à un environnement qui est vu de l'intérieur. Si un objet a une topologie en convexe étoilé, alors sa texture peut naturellement être enregistrée sous forme d'image omnidirectionnelle (Hernandez *et al.*, 2015). L'image représente alors son enveloppe extérieure, là où une image omnidirectionnelle d'un environnement représente une enveloppe intérieure.

3.1.3 Réalité Virtuelle 360°

Aujourd'hui, les artistes proposent des narrations en réalité virtuelle pour immerger totalement le spectateur dans leurs histoires. Les auteurs ont proposé le terme réalité virtuelle cinématique pour nommer ces applications, son contenu s'étendant de la vidéo immersive passive à la vidéo interactive où les choix du spectateur influencent le déroulé

de l'histoire (MacQuarrie et Steed, 2017). La réalité virtuelle cinématique se recoupe souvent avec la réalité virtuelle basée image (*Image-Based Virtual Reality*, *IBVR*) qui immerge l'utilisateur dans des scènes générées à partir d'images du monde réel (photographies ou vidéos) grâce aux méthodes du rendu basé image (Dupont de Dinechin, 2020). Les scènes de réalité virtuelle basée image peuvent être des représentations sans géométrie comme des images omnidirectionnelles ou des *light fields*, des représentations à géométrie implicite comme des images stratifiées, ou des représentations à géométrie explicite comme des grilles de voxels, des maillages ou des nuages de points (Richardt *et al.*, 2020; Chang et Wang, 2018). Les images et vidéos omnidirectionnelles sont alors les scènes les plus simples pour la réalité virtuelle cinématique et la réalité virtuelle basée image.

Un utilisateur de réalité virtuelle est immergé dans une image omnidirectionnelle de manière similaire indépendamment du dispositif et de son niveau d'immersion. Une image omnidirectionnelle étant la surface intérieure d'une sphère vue depuis son centre, l'approche standard est de la visualiser à travers une *skybox*. La *skybox* est une technique utilisée dans les moteurs graphiques pour simuler les éléments lointains d'une scène, généralement un ciel avec des nuages ou d'autres éléments d'environnements. La *skybox* consiste alors en une texture d'arrière-plan qui entour la scène et enveloppe l'ensemble de l'espace. Elle peut être vue comme un volume texturé, une sphère ou un cube, où l'utilisateur est confiné à l'intérieur. La texture de la *skybox* peut naturellement être une image omnidirectionnelle afin d'avoir une texture couvrant l'ensemble du volume. Comme une image omnidirectionnelle représente un environnement, la *skybox* est parfaitement adaptée pour l'afficher comme une vue globale des éléments lointains de la scène. La vision du volume en sphère est favorisée par la projection équirectangulaire et la vision du volume en cube par la projection *cubemap* (avec un patch de la *cubemap* pour chaque face du cube). Les représentations en sphère ou en cube sont toutefois équivalentes car la *skybox* vise à représenter les éléments lointains, sa taille est donc virtuellement infinie.

Mais l'utilisation d'une *skybox* pour représenter l'environnement de la scène n'est pas sans conséquences. Si l'utilisateur est dans une sphère de taille infinie, le rendu de la texture de la sphère est invariant. En effet, l'ordre de grandeur des mouvements de l'utilisateur étant limité par rapport au rayon de la sphère, aucun changement ne sera observable sur la texture de la sphère. Dans une *skybox*, l'utilisateur contrôle la direction dans laquelle il regarde l'image omnidirectionnelle mais ne se déplace pas dedans. Si l'image est visualisée sur un dispositif non-immersif comme un écran, l'utilisateur oriente la fenêtre de visionnage avec les éléments d'interaction traditionnelle comme la souris ou le clavier. Dans un casque de réalité virtuelle, la fenêtre de visionnage est contrôlée naturellement avec l'orientation de la tête de l'utilisateur. Enfin, dans

un CAVE, la vidéo omnidirectionnelle est généralement projetée en entier, l'utilisateur décide où regarder sans interactions particulières avec le dispositif. Selon le type de contenu, les interactions de sélection, manipulation et navigation, propres à la réalité virtuelle (Bowman *et al.*, 2001), peuvent donc ne pas être disponibles. L'interaction de navigation libre à 6 degrés de liberté (*6-DoF*), position et orientation, n'est possible que si la scène contient des informations visuelles sur plusieurs positions, ou si elle contient des informations géométriques (implicites ou explicites). La parallaxe de mouvement, le changement apparent de la position des objets lorsque l'utilisateur bouge la tête ou se déplace, n'est donc pas disponible uniquement avec une image omnidirectionnelle, la scène ne permet qu'une navigation à 3 degrés de liberté (3-DoF), uniquement l'orientation, ce qui limite la perception de la profondeur. Il est tout de même possible de reproduire un autre indice de profondeur, la stéréoscopie, grâce à la stéréo omnidirectionnelle. La stéréoscopie est la perception du relief résultant de la vision binoculaire, où chaque œil reçoit une même image légèrement décalée vers la gauche ou la droite. La stéréo omnidirectionnelle, ou ODS pour *Omnidirectional Stereo*, applique ce principe en utilisant une paire d'images omnidirectionnelles gauche-droite, souvent empilées verticalement en projection équirectangulaire (Richardt, 2020). La technique est inventée par (Ishiguro *et al.*, 1992) et les applications portaient initialement sur l'estimation de la géométrie. Aujourd'hui, le format ODS est largement utilisé pour la réalité virtuelle 360°. La plupart des caméras omnidirectionnelles non-SVP sont capables de capturer le format, et les moteurs 3D de le rendre. Le rendu l'ODS est simple, une *skybox* par œil est utilisée pour afficher les images gauche et droite dans la même direction. La disparité dans les deux images omnidirectionnelle crée l'effet stéréoscopique. Il a été aussi observé que l'absence de parallaxe pouvait favoriser la cybermaladie avec des symptômes similaires au mal des transports comme la fatigue oculaire, la désorientation ou la nausée (LaViola, 2000). La présence de stéréoscopie ne suffit pas à diminuer la cybermaladie dans une scène sous forme de vidéo omnidirectionnelle (Narciso *et al.*, 2019). Enfin, l'absence de parallaxe amène à un problème pour créer des scènes collaboratives. Dans une scène où tous les utilisateurs sont à la même position, il est plus difficile de trouver des incarnations pertinentes.

3.2 Télé-Immersion dans une Image 360°

Pour avoir le système de télé-immersion le plus simple possible, nous souhaitons nous appuyer sur uniquement les images omnidirectionnelles pour notre première version. Nous cherchons à déterminer si avec une caméra simple omnidirectionnelle, il est possible de créer un lieu commun que plusieurs visiteurs peuvent rejoindre simultanément avec des incarnations en avatars. Une caméra omnidirectionnelle statique est alors disposée sur un lieu d'intérêt et diffuse à des utilisateurs distants en direct la vidéo

omnidirectionnelle. Les visiteurs visualisent de manière immersive la vidéo et doivent se sentir comme s'ils étaient présents sur le lieu d'intérêt, et un avatar 3D est ajouté dans la scène pour chaque visiteur immergé dans la vidéo. Nous avons alors proposé une approche pour réaliser cette première version du système en nous basant uniquement sur des images omnidirectionnelles sans informations 3D additionnelles. Puisque notre système suppose une caméra omnidirectionnelle statique, la solution proposée fonctionne de manière équivalente pour une image et une vidéo (en supposant que la vidéo soit synchronisée entre tous les utilisateurs). Idéalement, ce système peut servir de support pour l'enseignement à distance, où le professeur utilise la caméra omnidirectionnelle pour capturer la salle de classe.

3.2.1 Approches Existantes

Une règle standard dans le domaine du rendu 3D est que si deux utilisateurs se trouvent à deux positions différentes dans la scène, ils doivent avoir deux rendus différents. Cette règle est la simple reproduction du monde réel : si on change de point de vue, la perception du monde change. Cependant, pour respecter cette règle et être capable de créer des points de vue pour chacune des différentes positions, il est nécessaire de connaître complètement la scène. Cette contrainte n'est pas respectée lorsque la scène n'est modélisée que par une unique image omnidirectionnelle, car seule la vue à la position de la caméra est connue. Il s'agit donc de créer un environnement collaboratif avec des incarnations à partir d'une image omnidirectionnelle. Les caméras omnidirectionnelles sont adaptées pour capturer un lieu comme s'il était vu à la première personne, mais pas adaptées pour capturer des données destinées à être visionnées. Avec uniquement une image omnidirectionnelle, les utilisateurs ne voient la scène que du point de vue de la caméra et sont donc tous virtuellement à la même position, rendant inadapté l'usage d'avatars. Un problème peu étudié dans la littérature est de savoir comment réaliser une scène de télé-immersion à partir d'une image omnidirectionnelle, et notamment comment représenter un autre utilisateur immergé. Ce problème est majeur pour que les utilisateurs soient conscients de la présence des uns des autres dans la scène.

Une solution naïve consiste à utiliser une *skybox* et représenter les avatars 3D des utilisateurs qui peuvent se déplacer librement dans la scène (figure 3.5 gauche). Le problème est que la *skybox* ne permet pas de créer d'effets de parallaxe sur l'image omnidirectionnelle, ce qui résulte en la sensation que les avatars sont détachés de l'environnement et flottent dans la scène (voir sous-section 4.1.3). (Nguyen *et al.*, 2017) proposent de considérer que les différents utilisateurs sont effectivement tous à la position de la caméra et de représenter les autres utilisateurs par des rectangles sur la vidéo omnidirectionnelle correspondant à leurs champs de vision. Les utilisateurs sont

1. `https://facebook.com/4/videos/380215617419883`

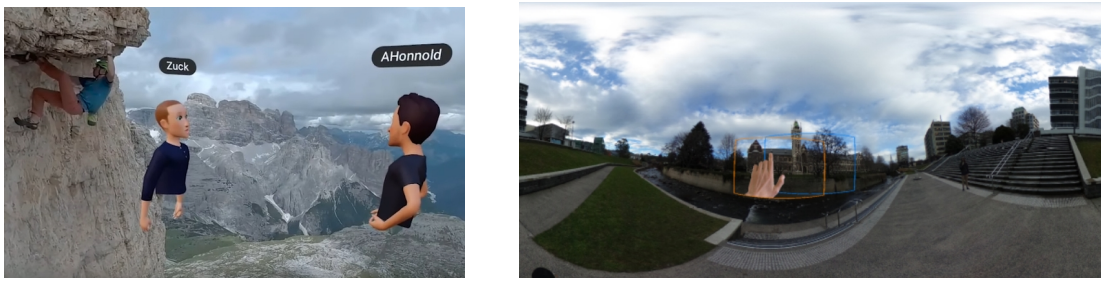


FIGURE 3.5 – Représentations des utilisateurs dans une vidéo omnidirectionnelle. Gauche : Vidéo omnidirectionnelle avec Meta Oculus Quest 2 (extrait vidéo [1]). Les utilisateurs sont représentés par des avatars 3D et peuvent se déplacer dans une *skybox*. Droite : (Young *et al.*, 2019). Les utilisateurs sont représentés par des rectangles correspondants à leurs champs de vision avec leurs mains.

alors conscients de la direction dans laquelle les autres utilisateurs regardent sans utiliser d'avatar. (Young *et al.*, 2019; Lee *et al.*, 2018) font une proposition similaire en représentant les utilisateurs par leurs champs de vision en ajoutant la main des utilisateurs en surimpression sur la vidéo (figure 3.5 droite). Bien que cette représentation soit plus efficace que de ne pas incarner les utilisateurs, la présence sociale est plus faible et la communication non-verbale plus limitée qu'avec des avatars. (Kumar *et al.*, 2022) proposent Tourgether360, un outil de visualisation collaborative pour la visite omnidirectionnelle de sites avec avatars. Dans cet outil, la caméra de la vidéo se déplace à l'intérieur du lieu avec les avatars des utilisateurs qui suivent la trajectoire de la caméra. La vidéo avec une caméra mobile permettant de ne plus être restreint à une unique position, les utilisateurs peuvent se déplacer le long du chemin de la caméra en avançant ou reculant dans la vidéo. Ainsi, des utilisateurs à des instants différents dans la vidéo peuvent se voir à travers leurs avatars. Même si cette approche fonctionne pour les lieux statiques, cette navigation spatio-temporelle pose un problème lorsque le lieu est dynamique. Par exemple, si les utilisateurs veulent inspecter un élément dynamique, celui-ci va changer d'état quand le point de vue à partir duquel il est observé change. Une revue avec de la cohérence entre des utilisateurs à différentes positions est alors impossible. Une autre solution pour une vidéo statique est d'utiliser des méthodes de reconstruction mono-vue, notamment les récentes approches issues de l'apprentissage profond (Fahim *et al.*, 2021), pour obtenir une scène 3D à partir d'une unique position de caméra. Avec une reconstruction complète du point de vue de la caméra, les utilisateurs peuvent se déplacer et interagir entre eux comme dans une scène 3D classique. Cependant, ce problème étant ambigu, la reconstruction est bruitée ou éloignée de la vérité terrain. Ces erreurs de reconstruction deviennent d'autant plus visibles que l'utilisateur s'éloigne de la position de la caméra. L'approfondissement de cette approche est détaillé au chapitre 4. Dans cette section, nous proposons de nous affranchir de la

FIGURE 3.6 – Ajout d'avatar dans une image omnidirectionnelle. Chacune des deux images correspond à une vue omnidirectionnelle d'un des deux utilisateurs. Dans les deux cas, l'image omnidirectionnelle est la même, mais l'avatar n'est pas placé au même endroit. La vue de l'avatar de l'autre utilisateur entraîne l'illusion que celui-ci voit la scène d'un point de vue différent.

règle du rendu 3D et de créer une scène de télé-immersion multi-utilisateurs basée uniquement sur une unique vue omnidirectionnelle en ajoutant des avatars 3D. L'ajout des avatars 3D est fait de manière cohérente afin que chaque utilisateur semble occuper une position différente dans la scène.

3.2.2 Approche Proposée

L'idée de base de cette approche pour créer une scène multi-utilisateurs est de donner l'illusion à chacun des utilisateurs que les autres ont un rendu différent de la scène. Cette illusion s'appuie sur l'utilisation d'un avatar pour tous les utilisateurs : la vue d'un utilisateur à une position différente de la sienne laisse penser que celui-ci voit effectivement la scène d'un autre point de vue. Pour fonctionner, la vidéo omnidirectionnelle est considérée comme un environnement, c'est-à-dire uniquement des éléments de champ éloigné qui seront toujours derrière les utilisateurs. La vidéo omnidirectionnelle ne devra contenir que des éléments d'arrière-plan. Un résultat est donné figure 3.6 d'une image omnidirectionnelle avec deux utilisateurs. On parle de position virtuelle pour désigner la position de l'avatar d'un utilisateur dans la scène de l'autre, c'est-à-dire la position à laquelle l'autre utilisateur pense qu'il est. Comparée à une reconstruction 3D, cette approche permet une meilleure qualité visuelle car les utilisateurs restent à la position de la caméra, évitant ainsi les artéfacts de reconstruction comme les surfaces fantômes. Pour arriver à ce résultat, les avatars doivent être placés de manière cohérente dans l'espace 3D. Cet espace 3D cohérent est obtenu en deux étapes. La première consiste à placer les utilisateurs autour de la position de la caméra dans une scène de référence. Ensuite, chaque utilisateur est immergé dans une copie de la scène de référence sur lequel une translation est appliquée afin que chacun des utilisateurs soit centré sur la position de la caméra. Un utilisateur appartient alors à une scène locale ayant sa position pour origine. Le principe est illustré figure 3.7. Pour que d'autres objets soient

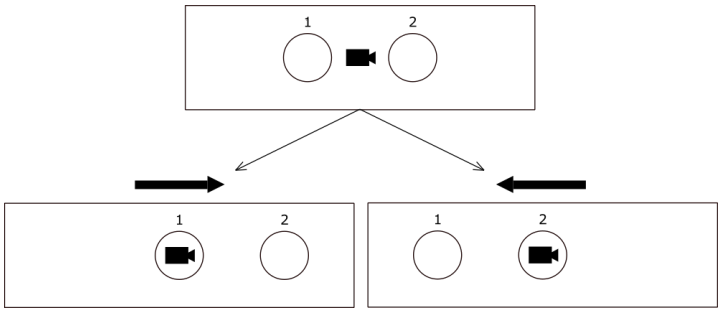


FIGURE 3.7 – Création des scènes locales centrées sur chacun des utilisateurs.

aussi ajoutés de manière cohérente à la scène, la translation de la scène locale doit aussi leur être appliquée.

Nous avons fait tester cette approche à quelques personnes lors de démonstrations. Les retours sont que ceux-ci n'avaient pas remarqué qu'ils visionnaient la même image omnidirectionnelle. La vue de l'autre utilisateur à une certaine position sur l'image omnidirectionnelle amène naturellement l'utilisateur à penser qu'il est réellement à cette position dans la scène. Cette approche présente donc un certain intérêt pour une application d'enseignement en ligne où enseignant donne cours sur un lieu filmé par une caméra omnidirectionnelle. Malheureusement, cette approche n'a pas été creusée car elle se confronte à certaines limitations dans l'immersion et l'interaction.

3.2.3 Limites

Pour cette première itération dans la conception d'un système de télé-immersion, nous avons souhaité utiliser les données les plus élémentaires disponibles avec une caméra omnidirectionnelle. Nous avons alors voulu rassembler des visiteurs dans des images et vidéos omnidirectionnelles sans informations supplémentaires. Bien que cette approche offre une solution facile à mettre en place, celle-ci se confronte à des problèmes sur la gestion des occultations et la mise en place d'interactions basiques.

Incohérence des Occultations

L'illusion que les deux utilisateurs sont en fait à la même position peut être brisée avec un type particulier de lieu Cette illusion reposant simplement sur l'intégration des avatars des différents utilisateurs à différentes positions dans une vidéo omnidirectionnelle, des incohérences dans les éléments visibles peuvent survenir. Ces incohérences sont la conséquence de la topologie en convexe étoilé capturée par la caméra omnidirectionnelle (sous-section 3.1.2). Si la topologie du lieu est convexe, l'illusion persiste car l'ensemble des éléments est visible depuis toutes les positions. Dans ce cas, il n'y a pas d'occultations, un utilisateur peut voir tous les éléments de la scène indépendamment

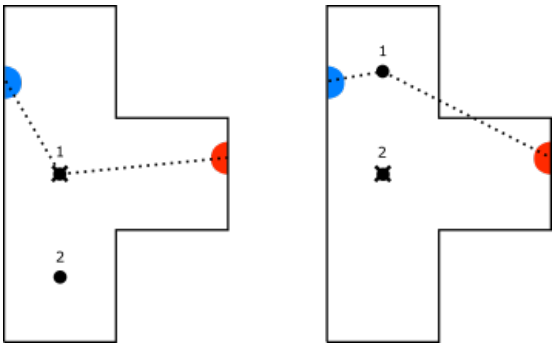


FIGURE 3.8 – Occultations incohérentes dans l'image omnidirectionnelle. Gauche : Scène locale centrée sur l'utilisateur 1 vue du dessus. La position de l'utilisateur étant celle de la caméra, les éléments bleu et rouge sont visibles. Droite : Scène locale centrée sur l'utilisateur 2 vue du dessus. Dans cette configuration, l'élément rouge n'est pas visible depuis la position de l'utilisateur 1. L'incohérence vient du fait que l'utilisateur 1 voit parfaitement l'élément rouge alors que l'utilisateur 2 pense alors qu'il ne le voit pas.

de sa position virtuelle, l'illusion reste donc valide. En revanche, si le lieu est un convexe étoilé, il est possible qu'il n'existe pas de chemin entre un élément de la vidéo omnidirectionnelle et la position virtuelle d'un utilisateur. Cet élément appartient à une région occultée et ne devrait donc pas être visible par l'utilisateur. Or, l'utilisateur étant en réalité à la position de la caméra, cet élément est de fait visible, même si sa position virtuelle semble indiquer le contraire. Comme représenté figure 3.8, un utilisateur peut alors voir des éléments dans sa scène qui devraient être occultés depuis sa position dans la scène de l'autre utilisateur. Cette illusion, laissant penser que l'autre utilisateur voit la scène d'un point de vue différent, peut alors être brisée si le lieu est un convexe étoilé. Toutefois, ce problème ne se présente pas si la scène représente un lieu extérieur où les éléments semblent être dans un horizon lointain. Les lieux extérieurs sont donc à favoriser avec cette approche.

Immersion et Interactions

En modélisant une scène de télé-immersion à partir d'un unique point de vue sans informations géométriques, des limites évidentes vont survenir pour l'immersion et l'interaction. La première limite apparente est l'absence de navigation. En effet, un seul point de vue étant connu et n'ayant pas d'information géométrique, il est impossible de voir le lieu sous une autre position. La gestion des occultations est aussi impossible sans géométrie car il est impossible de savoir si un élément de l'image omnidirectionnelle est devant ou derrière l'avatar. Un élément de premier plan sur l'image omnidirectionnelle ne peut pas être rendu devant un utilisateur.

Pour ajouter des fonctionnalités, nous avons mis en place une métaphore de pointeur laser comme dans les scènes de réalité virtuelle classiques. Ce pointeur permet d'ajouter

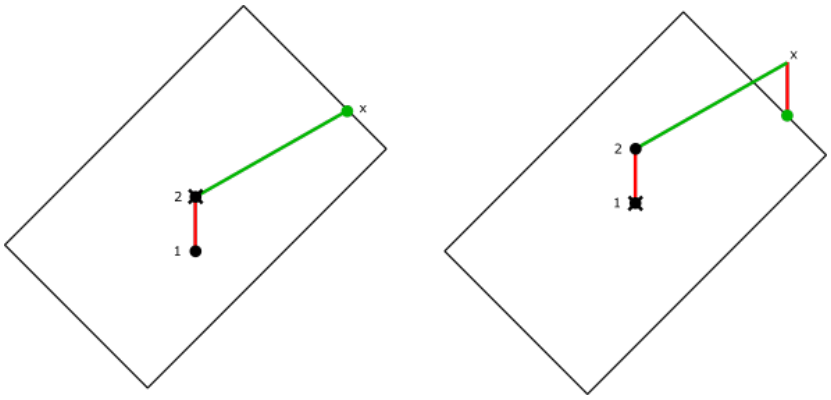


FIGURE 3.9 – Métaphore de pointeur laser et correction du point de contact. Gauche : Scène locale centrée sur l'utilisateur 2 vue du dessus. L'utilisateur 2 lance un laser sur l'environnement. Le laser intersecte l'environnement au point de contact x. Droite : Scène locale centrée sur l'utilisateur 1 vue du dessus. L'utilisateur 1 reçoit le point de contact x, mais comme l'origine du monde n'est pas la même, un décalage doit être appliqué. Ce décalage correspond à la translation entre les deux scènes.

de l'interaction entre les utilisateurs afin qu'ils puissent désigner des éléments de la vidéo omnidirectionnelle. En pratique, le laser doit partir de la main de l'avatar et s'arrêter au point de contact avec l'environnement. Cependant, comme l'origine de la scène dépend de l'utilisateur, ce point de contact n'est valable que dans la scène locale de l'utilisateur. Le besoin de translater le point de contact est illustré figure 3.9. Cette translation correspond au décalage entre les deux utilisateurs. Lorsqu'un utilisateur reçoit le point de contact ciblé par le laser d'un autre utilisateur, la translation est appliquée pour obtenir le point de contact rectifié et un laser est tracé entre cet autre avatar et ce point de contact. À noter que le décalage appliqué au laser fait qu'il ne fonctionne correctement que s'il est orienté vers un élément de la vidéo omnidirectionnelle. À cause du décalage entre les scènes locales, l'implantation d'un pointeur est impossible sans connaître le point de contact entre la scène et le laser. Une reconstruction 3D du lieu est alors impérative pour déterminer ce point de contact et l'interaction de pointage n'a pu être intégrée qu'en combinant l'image omnidirectionnelle avec des informations géométriques (chapitre 4).

À noter que cette reconstruction est réellement nécessaire lorsque la vidéo omnidirectionnelle représente des éléments qui sont à une distance du même ordre de grandeur que la distance entre les utilisateurs, comme dans un environnement en intérieur. Lorsque l'élément pointé paraît être dans un horizon lointain, par exemple si l'environnement représente un lieu extérieur, alors la longueur du laser peut être considérée comme beaucoup plus grande que la distance entre les utilisateurs qui peut être négligeable. Dans ce cas, le point de contact est commun aux deux scènes locales, la trajectoire du pointeur est retrouvée sans translation. La cohérence entre les scènes locales sera maintenue et

l'illusion aussi.

Ces limites dans l'immersion et l'interaction sont la conséquence du manque d'informations géométriques dans une image omnidirectionnelle. Certaines limites tiennent surtout pour les lieux intérieurs (occultations, décalages...) mais peuvent être ignorées si le lieu est un lieu extérieur Cette proposition de télé-immersion peut alors être intéressante pour un système où le lieu commun est en extérieur et que l'interaction de navigation libre n'est pas essentielle. Mais des améliorations doivent être apportées pour être un support de cours, qui sont généralement tenus en intérieur, surtout si l'on désire une navigation libre des étudiants dans la salle.

3.3 Conclusion

La caméra omnidirectionnelle est un choix raisonnable comme dispositif d'acquisition de télé-immersion. Les images omnidirectionnelles représentant une vue complète du lieu vu de la position de la caméra, elle permet de télé-immerger un utilisateur à cette position exacte. Cependant, ces données ne permettant de voir le lieu que d'un seul point de vue, la télé-immersion de plusieurs utilisateurs dans une image omnidirectionnelle n'est pas naturelle. Si les utilisateurs voient tous la scène depuis la position de la caméra, alors ils sont à la même position dans la scène et ne peuvent pas se percevoir entre-eux à travers des avatars. Nous avons alors proposé un système de télé-immersion qui rassemble dans une image omnidirectionnelle plusieurs utilisateurs incarnés en avatars pour améliorer la présence sociale. Ce système fonctionne en donnant l'illusion que les utilisateurs sont à des positions différentes et est parfaitement adapté pour télé-immerger des utilisateurs dans des lieux extérieurs si le cas d'usage ne requiert pas qu'ils puissent se déplacer librement. L'utilisation d'un seul point de vue limite tout de même les possibilités d'immersion et d'interaction dans les lieux intérieurs, la principale restriction étant l'impossibilité de pouvoir contrôler librement sa position dans la scène. La proposition doit alors être améliorée pour servir de support d'enseignement à distance avec des élèves se déplaçant librement dans la salle de classe.

Nous allons donc dans la suite explorer les approches pour augmenter les images omnidirectionnelles d'informations géométriques pour télé-immerger des visiteurs dans des lieux intérieurs et mettre en place une navigation libre dans la scène.

Chapitre 4

Télé-Immersion 3D 360° Statique

Dans le chapitre précédent, nous avons exploré la télé-immersion de plusieurs utilisateurs dans des simples images et vidéos 360°. Cependant, nos expérimentations sur la conception de scènes à partir de ces données nous ont conduits à la conclusion que le manque d'informations 3D limitait l'immersion et l'interaction. Dans ce chapitre, nous allons présenter des solutions pour augmenter les images 360° d'informations 3D afin d'obtenir un système de télé-immersion plus immersif et interactif. Pour simplifier le problème, nous nous intéresserons uniquement aux lieux statiques et utiliserons une approche hors-ligne qui peut ne pas être compatible pour atteindre un système temps réel. Pour réaliser ce nouveau système de télé-immersion, nous explorons l'utilisation conjointe de l'image 360° avec la carte de profondeur. Dans un premier temps, nous étudions possibilités offertes par cette modélisation avec une image 360° et sa carte de profondeur, et comment l'obtenir à partir de données réelles. Nous utilisons ensuite ces possibilités d'immersion et d'interaction pour télé-immerger des étudiants dans des environnements 3D de lieux d'intérêt préalablement reconstruits.

4.1 Image 360° et Géométrie

Précédemment, nous avons télé-immergé des utilisateurs sur des lieux en manipulant des images omnidirectionnelles simples. Cependant, les limites de l'immersion et de l'interaction avec ces données, nous ont amené à envisager l'ajout d'informations, notamment géométriques. Cet ajout de géométrie vise dans un premier temps à atteindre un objectif bien précis : se déplacer dans une image omnidirectionnelle. Dans cette section, nous présenterons les nouvelles fonctionnalités implantées en enrichissant l'image omnidirectionnelle d'informations géométriques sous forme de carte de profondeur. Nous discuterons aussi de comment estimer ces informations géométriques et des limites de la représentation.

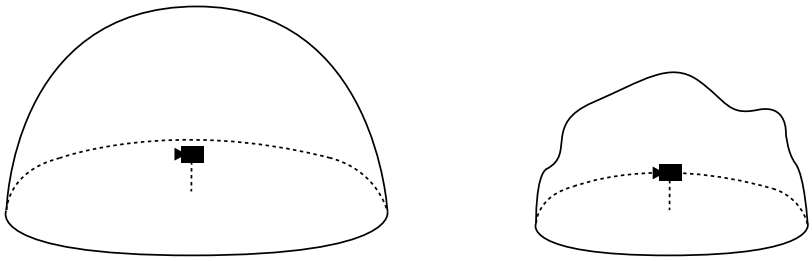


FIGURE 4.1 – Visitation d'image omnidirectionnelle avec et sans proxy géométrique. Gauche : Visualisation sans proxy géométrique (*skybox*). L'image est projetée sur la surface intérieure d'une sphère de rayon infini. Droite : Visualisation avec proxy géométrique. L'image est projetée sur la surface intérieure d'un volume quelconque de taille finie.

4.1.1 Proxy Géométrique

Sans informations géométriques, l'image omnidirectionnelle qui compose la scène est rendue à travers une *skybox*. L'utilisateur est à l'intérieur de l'image et peut regarder dans toutes les directions, mais il ne peut pas se déplacer à l'intérieur de la scène. La visualisation en *skybox* peut alors être considérée comme la projection de l'image sur une sphère de rayon arbitrairement grand, l'utilisateur se trouvant à l'intérieur, au centre de la sphère. Comme l'échelle de distance dans laquelle l'utilisateur se déplace (par exemple lorsqu'il bouge la tête) est trop petite par rapport au rayon de la sphère, le mouvement n'a aucun impact sur le rendu de l'image. Pour intégrer l'interaction de navigation dans une image omnidirectionnelle, le mouvement de l'utilisateur doit entraîner des changements cohérents dans l'affichage de l'image. Une solution est alors d'utiliser un proxy géométrique (Bertel *et al.*, 2019, 2020). L'idée est de projeter plutôt l'image omnidirectionnelle sur un volume, qu'on appelle proxy géométrique, dont la dimension est du même ordre de grandeur que l'échelle des mouvements de l'utilisateur. À l'instar d'une *skybox*, l'utilisateur est placé à l'intérieur du volume et l'image est projetée sur la surface intérieure. Toutefois, du fait que la taille du proxy est du même ordre de grandeur que les déplacements de l'utilisateur, ses mouvements ont un effet direct sur le rendu. Le proxy rend ainsi possible la visualisation de l'image omnidirectionnelle avec une navigation à l'intérieur. Le proxy se heurte tout de même à la limite qu'il ne peut pas reproduire les effets dépendants de la vue comme les reflets de lumière. Dans ce chapitre, la reconstruction 3D d'un lieu à partir d'une seule image omnidirectionnelle consiste donc à trouver un proxy géométrique.

Le proxy peut être une figure élémentaire telle qu'une sphère de rayon fini, ou un hémisphère pour projeter le bas de l'image sur un plan au sol et le haut sur une moitié de sphère. Cependant, une figure 3D plus complexe adaptée à l'image omnidirectionnelle est souhaitable pour reproduire les profondeurs et éviter les distorsions. En effet,

le proxy apportant du relief à l'image, il convient que les éléments qui semblent proches soient projetés sur les régions saillantes du proxy et que les éléments distants soient projetés sur les régions enfoncées. Cette idée de projeter l'image sur un proxy est similaire à celle de la réalité augmentée par projection avec un projecteur qui texture une surface vierge (Mamone *et al.*, 2021), ou à la création d'illusions anamorphiques. Ces illusions correspondent aux projections déformées visibles uniquement d'un point de vue spécifique, comme le dessin d'un faux trou sur le sol. Lorsqu'un artiste veut créer cette illusion, il réalise un dessin donnant un effet de perspective en fonction de la surface. Avec le rendu par proxy, le processus est inversé : la texture est connue (l'image omnidirectionnelle) et l'objectif est de trouver la surface qui donne l'illusion adéquate lorsque l'image y est projetée. Le choix du proxy est essentiel afin que l'illusion créée ne se dégrade pas excessivement lorsque l'utilisateur s'éloigne de la position de la caméra. Bien que la forme du proxy puisse être quelconque, l'ajout de restrictions topologiques au proxy est souhaitable en raison des propriétés de la caméra omnidirectionnelle. En effet, le système d'acquisition ne capture pas toutes les informations visuelles du lieu mais seulement celles qui sont directement perceptibles, la caméra capture la surface d'un convexe étoilé. (sous-section 3.1.2). Par conséquent, si l'image est projetée sur un proxy qui n'est pas un convexe étoilé (tel qu'une forme en U), certaines régions de la surface du proxy seront dépourvues de texture. Un proxy en convexe étoilé est donc nécessaire pour obtenir une scène 3D entièrement texturée. Face à ces contraintes, la carte de profondeur omnidirectionnelle s'affirme comme un choix judicieux pour modéliser un proxy géométrique.

4.1.2 Carte de Profondeur Omnidirectionnelle

Traditionnellement, une carte de profondeur est une image en niveau de gris tel qu'un pixel encode la distance entre la caméra et ce pixel. Cette donnée permet représenter la géométrie sous forme d'image. Son équivalent omnidirectionnel est aussi une image en niveaux de gris, mais un pixel représente la distance entre ce point particulier de la surface de la sphère et le centre de la sphère (Dupont de Dinechin et Paljic, 2018). La donnée conjointe d'une image couleur avec une carte de profondeur est appelée image RGB-D (Attal *et al.*, 2020), vue plus profondeur (V+D) (Azzari *et al.*, 2010), couleur plus profondeur (da Silveira et Jung, 2019) ou panorama monoscopique plus profondeur pour les images 360° (Dupont de Dinechin et Paljic, 2018). Cette représentation est aussi parfois qualifiée de 2.5D (Dhamo *et al.*, 2019) car entre l'image 2D et la reconstruction 3D complète. Les mêmes projections que pour une image omnidirectionnelle peuvent être utilisées pour la carte de profondeur (équirectangulaire, *cubemap*...), et celle-ci peut être enregistrée selon plusieurs encodages : z-distance, disparité ou disparité normalisée (Wegner *et al.*, 2017). L'encodage z-distance étant le plus simple à manipuler, c'est

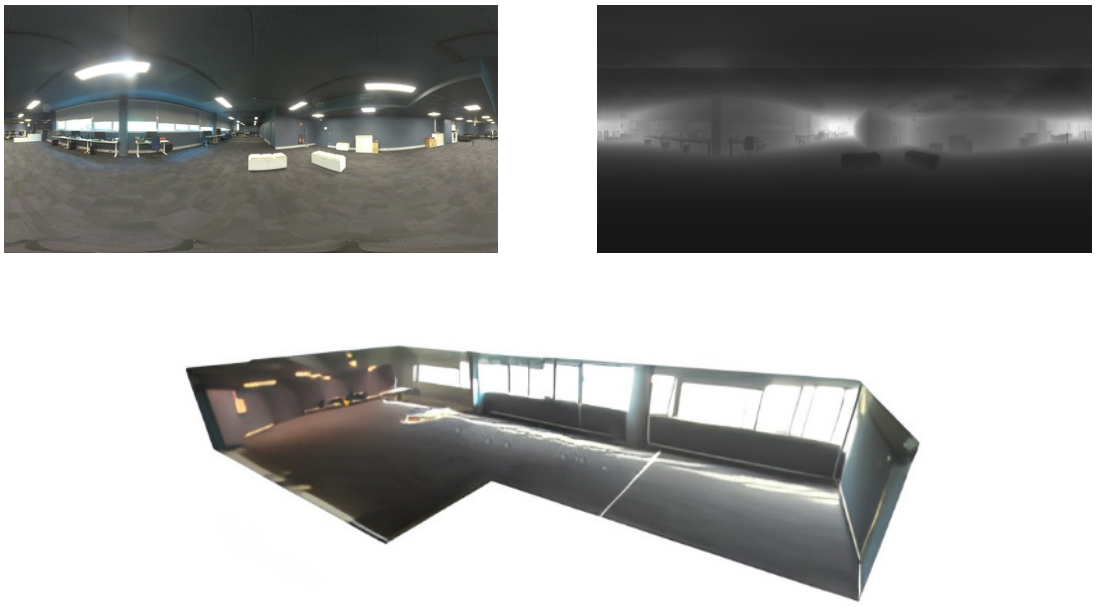


FIGURE 4.2 – Image omnidirectionnelle avec géométrie. Haut-Gauche : Image omnidirectionnelle. Haut-Droite : Carte de profondeur omnidirectionnelle. Bas : Reconstruction 3D.

avec ce format que nous avons reconstruit nos représentations 3D. Un exemple de carte de profondeur avec cet encodage est donné figure 4.2 en haut à droite (avec comme convention que les pixels les plus sombres sont les plus proches de la caméra et les pixels clairs les plus éloignés). À l'aide de données numériques supplémentaires, la carte de profondeur permet d'obtenir une reconstruction 3D du lieu à l'échelle dans laquelle un utilisateur peut être immergé.

Par définition, une carte de profondeur omnidirectionnelle encode en 3D une sphère radialement déformée. La carte de profondeur étant une discrétisation de la fonction de rayon (sous-section 3.1.2), cette déformation radiale résulte toujours en une topologie en convexe étoilé. Grâce à ces propriétés, la carte de profondeur omnidirectionnelle se convertit naturellement en un nuage de points où chaque pixel devient en point 3D. Avec la projection équirectangulaire, un pixel (i, j) de la carte de profondeur ayant pour valeur d est converti en un point 3D (x, y, z) avec la formule suivante :

$$\begin{aligned} x &= s\, d \cos\theta \sin\varphi \\ y &= s\, d \cos\varphi \\ z &= s\, d \sin\theta \sin\varphi \end{aligned} \tag{4.1}$$

où (θ, φ) est la conversion des coordonnées du pixel (i, j) en leurs angles de latitude et longitude (calculés avec l'équation 3.1), et s est la valeur d'échelle pour passer de la

FIGURE 4.3 – Ajout d'un objet dans une image omnidirectionnelle sans et avec occultations. Gauche : Ajout naïf dans une image omnidirectionnelle. Sans carte de profondeur, l'objet est ajouté au premier plan devant les éléments de l'image omnidirectionnelle. Cette approche crée des problèmes d'occultations si des éléments de l'image omnidirectionnelle doivent être devants l'objet. Droite : Ajout d'un objet dans une image omnidirectionnelle avec la carte de profondeur. Les informations de profondeur permettent de savoir quelles régions de l'objet doivent être dissimulées.

valeur de profondeur d à la distance réelle. Cette valeur s est généralement la distance réelle du pixel le plus éloigné sur la carte de profondeur quand d est normalisée entre 0 et 1. La couleur du pixel (i, j) sur l'image omnidirectionnelle est attribuée au point (x, y, z). Ce nuage de points texturé est facilement rendu avec la plupart des moteurs graphiques mais le problème est que le rendu sera épars, en particulier pour les points avec des valeurs de profondeur élevées qui sont éloignés les uns des autres. Pour un rendu dense, la solution consiste à créer un maillage à partir de la carte de profondeur. Ce maillage est créé en s'appuyant sur l'hypothèse que la reconstruction est un convexe étoilé, et donc qu'il existe toujours une surface entre des points voisins. Cette hypothèse mène naturellement à créer des surfaces triangulaires entre les coordonnées 3D de 3 pixels voisins dans la carte de profondeur (Dziembowski *et al.*, 2016). Un exemple de reconstruction est illustré figure 4.2 en bas.

4.1.3 Immersion et Interaction

Comme présenté au chapitre précédent, une scène avec seulement image omnidirectionnelle sans informations géométriques est limitée en immersion et en interaction. La visualisation en *skybox* ne permet pas de restituer des indices sur la profondeur de la scène comme les occultations ou la parallaxe de mouvement. L'introduction d'une carte de profondeur adaptée à l'image permet de résoudre ces problèmes simplement en utilisant la reconstruction 3D.

Le manque d'occultation avec une *skybox* impose que des objets 3D ajoutés dans la scène ne seront pas masqués par des éléments de l'image omnidirectionnelle et seront

FIGURE 4.4 – Génération de parallaxe à partir d'une image omnidirectionnelle avec la carte de profondeur associée. Gauche : Image omnidirectionnelle originale. Droite : Parallaxe vers la droite.

donc toujours des éléments de premier plan. La reconstruction 3D avec les règles de rendu classique affichant un élément devant un autre si celui-ci est bien devant dans l'espace 3D résout ce problème (figure 4.3).

La reconstruction 3D permet aussi naturellement de créer la parallaxe de mouvement comme illustré figure 4.4. Toutefois, la génération de parallaxe avec cette méthode de reconstruction montre quelques inconvénients. Un premier inconvénient est qu'en se déplaçant de la position de la caméra, des surfaces fantômes vont devenir visibles, conséquence du fait que la reconstruction 3D connecte des surfaces qui sont en réalité indépendantes. Ces artéfacts en forme de surfaces étirées dégradent la qualité du rendu quand on s'éloigne de la position de la caméra. Un second inconvénient est que, la résolution de l'image omnidirectionnelle étant limitée, le rendu devient de plus en plus flou à mesure que l'on s'éloigne de la position de la caméra. Les points 3D les plus éloignés sont étalés sur une surface plus large pour couvrir les trous et obtenir un rendu dense.

Même si l'effet de parallaxe présente quelques inconvénients, il résout un problème majeur de la représentation avec uniquement une image omnidirectionnelle. Lorsque des objets 3D sont ajoutés dans une scène modélisée par une image omnidirectionnelle, la visualisation de l'image en *skybox* donne l'impression que ceux-ci flottent. En effet, en l'absence de parallaxe dans l'image omnidirectionnelle, les objets ajoutés dans la scène semblent ne pas être ancrés dans la scène, le mouvement autour d'un objet entraîne un changement de rendu de celui-ci alors que l'image omnidirectionnelle reste la même comme si les éléments de l'image étaient éloignés. L'effet de parallaxe existe pour l'objet 3D mais pas pour l'image omnidirectionnelle, ce qui donne l'impression que l'objet flotte. Cet effet peut n'est pas dérangeant si l'image omnidirectionnelle représente des éléments perçus comme lointains (par exemple la scène reste crédible si l'objet flotte dans une image de l'espace interstellaire). Mais cela peut être incohérent dans le cas où les éléments de l'image sont interprétés comme proches comme dans un environnement intérieur (figure 4.5 haut). La parallaxe créée par la reconstruction 3D change de

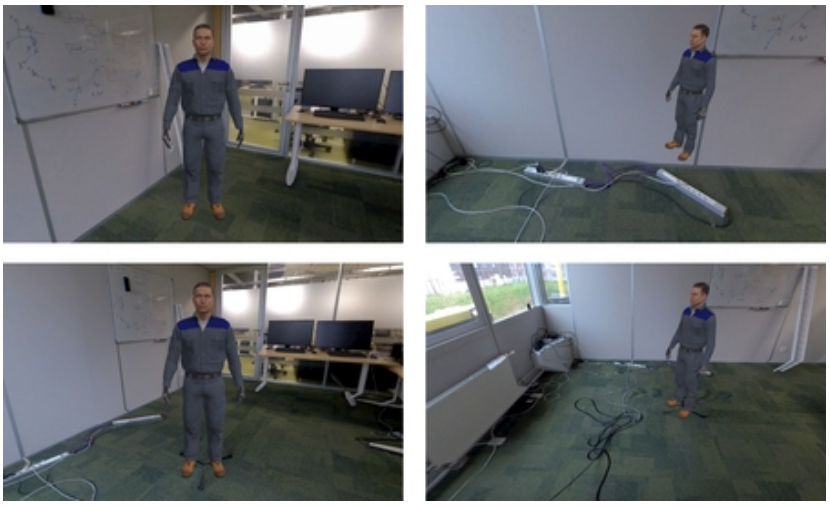


FIGURE 4.5 – Effet de flottement des objets dans une image omnidirectionnelle. Haut : Rendu de l'image omnidirectionnelle sous forme de *skybox*. L'objet flotte dans la scène. Bas : Rendu de l'image omnidirectionnelle sous forme de reconstruction 3D. L'objet est bien incrusté dans la scène.

manière cohérente le rendu des objets 3D ajoutés dans la scène et le rendu de l'image omnidirectionnelle, incrustant bien ainsi un objet dans la scène.

Enfin, la carte de profondeur permet aussi d'estimer la normale en un point. Cette information est importante pour l'immersion, car elle permet d'augmenter la qualité du rendu avec des effets de lumière, mais également pour l'interaction comme la gestion des collisions. La première étape pour estimer la normale d'un pixel (i, j) sur la carte de profondeur est de calculer une projection gnomonique autour de (i, j) pour obtenir un voisinage carré sans distorsions. Les coordonnées 3D des points du voisinage sont ensuite calculées à l'aide des valeurs de profondeur. Un plan 3D passant par cet ensemble de points 3D est alors approximé, la normale du plan correspond à la normale du pixel (i, j). Cet algorithme n'est pas parfaitement robuste car l'approximation du plan peut être sensible aux valeurs aberrantes de la carte de profondeur, mais une estimation avec corrélation augmente la fiabilité sur des données bruitées (Du *et al.*, 2020).

4.1.4 Estimation de la Carte de Profondeur

La carte de profondeur est une information intéressante pour la création de scènes de télé-immersion plus immersives et interactives. Récupérer la carte de profondeur à partir de données réelles est donc essentiel pour améliorer notre système de télé-immersion 360° dans des lieux réels. Néanmoins, l'estimation de cartes de profondeur à partir d'une caméra omnidirectionnelle reste un problème de recherche actif. Nous avons étudié les moyens pour l'obtenir à partir de données pouvant être capturées avec une caméra omnidirectionnelle statique.

FIGURE 4.6 – Approches pour l'estimation de profondeur. Les images correspondent aux reconstructions 3D obtenues avec les différentes cartes de profondeur. Gauche : Estimation directe de la carte de profondeur. Droite : Estimation de l'agencement.

Approches Existantes

Un premier type d'approche pour obtenir une carte de profondeur à partir d'une seule caméra statique consiste à récupérer la géométrie des éléments visibles en utilisant des méthodes de reconstruction mono-vue (Fahim *et al.*, 2021). Avec le développement des réseaux de neurones, il est aujourd'hui possible d'estimer une carte de profondeur simplement à partir d'une image omnidirectionnelle. Ces réseaux peuvent être catégorisés en deux types d'approches : l'estimation directe de la carte de profondeur et l'estimation de l'agencement (da Silveira *et al.*, 2022). Un exemple de reconstruction 3D produite avec ces deux approches est donné figure 4.6. Le premier type de méthode vise à calculer la carte de profondeur directement en évaluant la distance à la caméra pour chaque pixel. Différents auteurs ont proposé des réseaux de neurones produisant une carte de profondeur omnidirectionnelle, en prenant comme entrée une projection équirectangulaire de l'image (Sun *et al.*, 2021; Zioulis *et al.*, 2018), un découpage en images perspectives (Li *et al.*, 2022; Rey-Area *et al.*, 2022), ou à la fois des images équirectangulaires et perspectives (Jiang *et al.*, 2021; Wang *et al.*, 2020). Le découpage en images perspectives est avantageux car il permet d'exploiter les architectures d'estimation de profondeur monoculaire déjà existantes pour les images classiques. Les cartes de profondeur des différentes images sont agrégés en une carte de profondeur omnidirectionnelle globale. Ces méthodes d'estimation directe tendent à produire une géométrie plus fine, mais le problème est difficile à résoudre en raison des nombreuses ambiguïtés. Les erreurs d'estimation avec cette approche entraînent des artéfacts qui dégradent l'immersion. De leurs côtés, les méthodes d'estimation de l'agencement permettent de retrouver une carte de profondeur simplifiée d'un lieu intérieur en approximant le sol, le plafond et les murs comme des plans. Celles-ci détectent des caractéristiques dans l'image, telles que les

coins de la pièce, et se servent d'hypothèses communes sur les lieux intérieurs pour retrouver une représentation géométrique complète. Une hypothèse fréquemment utilisée pour l'estimation d'agencement est l'hypothèse de Manhattan (Wang *et al.*, 2021a; Sun *et al.*, 2019; Yang *et al.*, 2019) qui suppose que les murs du lieu s'intersectent en angle droit. Bien que ces approches soient plus robustes, elles présentent l'inconvénient de produire des cartes de profondeur où les éléments en relief sont aplatis contre les murs, les sols et les plafonds. En conséquence, les scènes de réalité virtuelle qui en découlent sont moins réalistes. Ces deux approches peuvent être exploitées conjointement pour l'estimation de profondeur d'un lieu intérieur avec une approche mixte (Zeng *et al.*, 2020). D'autres approches alternatives ont été proposées, reposant sur des données plus complexes qu'une simple image omnidirectionnelle, afin d'exploiter les propriétés des appareils multi-caméras. Par exemple, (Meuleman *et al.*, 2021) ont utilisé une caméra omnidirectionnelle à quatre objectifs *fisheyes* et ont proposé de calculer la carte de profondeur en temps réel grâce à la stéréoscopie. Les dispositifs multi-caméras peuvent également capturer des images sous le format standard stéréo omnidirectionnelle ODS. Avec ces appareils, une carte de profondeur peut être déduite en calculant la disparité entre l'image gauche et l'image droite de l'ODS. Dans (Lai *et al.*, 2019), un réseau de neurones léger est proposé pour estimer la carte de profondeur à partir d'une image ODS en temps réel pour une approximation de la géométrie en direct.

Malgré ces nombreuses méthodes d'estimation de profondeur, aucune ne semble donner directement une carte de profondeur satisfaisante pour la réalité virtuelle avec nos données. Les reconstructions 3D des méthodes orientées estimation directe apparaissaient trop bruitées pour être crédibles en réalité virtuelle et les reconstructions des méthodes orientées estimation d'agencement sont erronées si le lieu est encombré, avec des erreurs dans la géométrie donnant des conflits accommodation-vergence à l'utilisateur (Ozkan et Celikcan, 2023).

Approche Proposée

Nous avons adopté une approche personnalisée simple, similaire à 360MonoDepth (Rey-Area *et al.*, 2022), en modifiant l'échantillonnage en images perspectives, le modèle d'estimation de la profondeur et la méthode de fusion des profondeurs. L'algorithme proposé est illustré figure 4.7. Pour l'échantillonnage, des images tangentes avec chevauchement sont extraites par projection gnomonique de façon à couvrir l'ensemble de la sphère. Nous avons décidé empiriquement d'utiliser un découpage de l'image sphérique en 15, 60 et 15 images perspectives respectivement aux valeurs de latitudes de $\frac{5\pi}{12}$, 0 et $-\frac{5\pi}{12}$ afin d'avoir une couverture fine de la géométrie pour un temps de calcul raisonnable (mais pas forcément temps réel). Le champ de vision est fixé à $\frac{\pi}{2}$ pour toutes les images afin de maximiser la couverture et réduire les distorsions. Pour l'estimation de

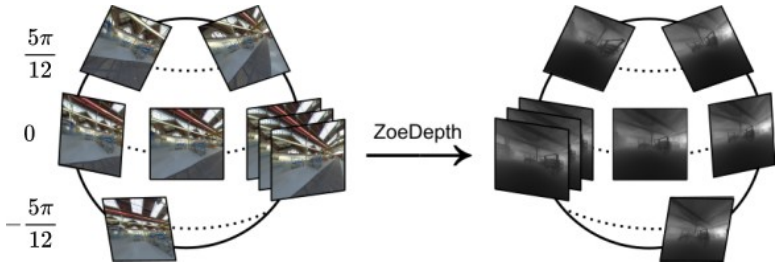


FIGURE 4.7 – Approche proposée pour l'estimation omnidirectionnelle de profondeur. L'image équirectangulaire d'entrée est décomposée en images perspectives. L'estimation monoculaire de la profondeur est effectuée sur chaque image perspective. Les cartes de profondeur résultantes sont projetées sur une image équirectangulaire commune.

profondeur, nous avons choisi d'utiliser le réseau ZoeDepth (Bhat *et al.*, 2023) au lieu de Midas (Ranftl *et al.*, 2020). Ce réseau prédit une profondeur métrique à partir d'une image perspective, c'est-à-dire une profondeur interprétée en unité physique (en mètres) au lieu d'une profondeur relative entre les éléments. La cohérence entre les différentes prédictions pour la fusion des profondeurs est grandement améliorée car les problèmes d'échelle entre les différentes cartes de profondeur sont évités. ZoeDepth est exécuté sur toutes les images perspectives et les cartes de profondeur sont projetées dans une carte de profondeur équirectangulaire globale à l'aide d'une projection gnomonique inverse. Les valeurs de profondeur se chevauchant sur plusieurs images sont fusionnées par une moyenne sur l'ensemble des cartes de profondeur. Chaque pixel étant couvert par plusieurs images, la carte de profondeur équirectangulaire produite est lisse et cohérente avec moins de distorsions et d'artéfacts. Cette méthode d'estimation semble empiriquement donner des reconstructions 3D réalistes pour la réalité virtuelle. Elle est une des bases pour la création de scènes dynamiques au chapitre suivant.

Une dernière difficulté pour la reconstruction 3D à partir d'images omnidirectionnelles réelles est que la carte de profondeur doit être accompagné de la valeur d'échelle s (équation 4.1). Sans cette valeur, l'utilisateur est immergé dans une scène qui n'est pas aux bonnes dimensions comparé au lieu réel. Une solution naïve est de supposer qu'une distance dans la carte de profondeur est connue a priori et d'estimer la valeur d'échelle en fonction de cette distance. Une distance intéressante à utiliser est la hauteur de la caméra omnidirectionnelle qui permet de retrouver s grâce aux valeurs de profondeur d'un pixel du bas de l'image omnidirectionnelle correspondant au sol. Cette méthode est choisie de préférence, car cette hauteur est nécessairement connue pour obtenir les scènes 3D au prochain chapitre.

4.1.5 Limites

L'image omnidirectionnelle avec carte de profondeur constitue une amélioration comparé à l'image omnidirectionnelle seule sur le plan de l'immersion et l'interaction. Le chapitre précédent a montré que l'image omnidirectionnelle seule ne permettait pas certaines interactions si l'image représente un lieu extérieur (comme le pointage laser ici possible). En combinant l'image à une carte de profondeur, les interactions qui étaient impossibles au chapitre précédent sur des lieux intérieurs sont désormais réalisables. De nouvelles fonctionnalités peuvent aussi être déployées comme les occultations ou la navigation libre. Néanmoins, la représentation de lieux extérieurs avec une carte de profondeur n'est pas complétement satisfaisante. En pratique, les images omnidirectionnelles d'extérieurs contiennent des éléments en champ proche (sol, personne...) et des éléments en champ lointain (horizon, ciel...). Le rendu par *skybox* est adéquat pour le champ lointain alors que le rendu par proxy est plus adéquat pour le champ proche. Avec la carte de profondeur, les éléments lointains peuvent tout de même être projetés à une distance arbitrairement large, mais comme tous les éléments doivent être connectés, des artéfacts vont apparaitre aux régions avec de grandes discontinuités de profondeur. Cette observation est aussi valable pour des lieux intérieurs complexes où des surfaces vont potentiellement connecter des éléments de l'arrière-plan avec des éléments du premier plan. Ainsi, comme la reconstruction considère que l'image représente une seule et même grande surface (une seule composante connexe), il n'est pas possible de sélectionner individuellement un élément de l'image, ou de tourner autour dans la scène 3D. Pour rendre des éléments indépendants, une option consiste à ne pas former de surface, par exemple si la distance entre deux éléments voisins dans la carte de profondeur dépasse un seuil prédéfini. Le problème est que ceci conduit à un rendu avec des trous, qui n'est pas forcément plus agréable visuellement. Fondamentalement, l'image omnidirectionnelle avec la carte de profondeur reste une information qui modélise l'environnement de la scène et n'est pas appropriée pour modéliser les objets. L'image omnidirectionnelle avec carte de profondeur reste donc à privilégier pour la télé-immersion dans des lieux intérieurs sans ou avec peu d'éléments de premier plan.

Enfin, pour télé-immerger des utilisateurs sur des lieux dynamiques, la carte de profondeur doit être étendue aux vidéos. Cette extension de la carte de profondeur aux vidéos omnidirectionnelle avec une caméra statique est faite de deux manières : soit en considérant une carte de profondeur par image de la vidéo (géométrie dynamique) soit en utilisant une même carte de profondeur pour toutes les images (géométrie statique). L'intérêt d'une même carte de profondeur pour toutes les images est restreint car la géométrie des éléments devra être la même le long de la vidéo et n'est adéquat que pour s'immerger dans des scènes avec des effets animés comme des effets de lumière. Associer une carte de profondeur pour chaque image de la vidéo est plus pertinent comme les

éléments mobiles de la scène pourront aussi être rendus en relief. Mais l'estimation de profondeur est une opération coûteuse en temps et nous n'avons pas trouvé ni réussi à développer de méthode pouvant la restituer en temps réel avec une qualité suffisante pour le visionnage en réalité virtuelle. À l'avenir, nous pouvons espérer que des capteurs de profondeur omnidirectionnels soient développés (Meuleman *et al.*, 2021; Son *et al.*, 2019; Zhou *et al.*, 2013) et que ces appareils se généralisent.

4.2 Télé-Immersion dans un Scan 3D

Dans la section précédente, nous avons introduit la carte de profondeur omnidirectionnelle et mis en évidence ses avantages et inconvénients. Même si cette représentation n'offre pas une navigation libre totalement satisfaisante, nous avons voulu utiliser ce type de données pour améliorer le système de télé-immersion précédent. L'idée est de télé-immerger des visiteurs sur un lieu capturé par une caméra omnidirectionnelle en utilisant les possibilités offertes par la carte de profondeur pour l'immersion et l'interaction. Mais comme détaillé précédemment, les images omnidirectionnelles avec cartes de profondeur ne permettent de modéliser que des environnements et non des objets. Pour immerger des visiteurs sur un lieu commun avec des objets, nous avons alors choisi de les réunir sur un lieu virtuel où les objets sont modélisés, un scan 3D. Les scans 3D, souvent stockés sous forme de nuages de points volumineux, sont intéressants car ils représentent des lieux où les informations de textures et de géométries sont entièrement connues. L'intérêt est qu'en connaissant un modèle complet du lieu, il est facile de générer des vues omnidirectionnelles avec cartes de profondeur pour chacune des positions. Les nuages de points sont communément utilisés en télé-immersion (Lee *et al.*, 2021; Gao *et al.*, 2017), notamment pour transmettre des représentations d'objets comme des avatars (Gamelin *et al.*, 2021; Yu *et al.*, 2021). Nous avons alors développé une application pour télé-immerger des étudiants à distance sur le scan d'un lieu d'intérêt qu'ils pourraient explorer librement. Cette application permet d'explorer uniquement des lieux sous forme de nuages de points statiques afin de nous concentrer sur l'étude des images omnidirectionnelles et non des vidéos. Dans ces nuages, le nombre de points reste constant et leurs positions restent fixes.

Aujourd'hui, de nombreuses techniques ont été développées pour acquérir l'apparence visuelle de lieux en nuage de points. Des dispositifs tels que les LiDAR (Raj *et al.*, 2020) ou les caméras RGB-D (Zollhöfer *et al.*, 2018) capturent directement sous cette représentation, mais aussi des simples caméras couleur lorsqu'elles sont combinées à des algorithmes tels que Structure-from-Motion et Multi-View-Stereo (Richardt *et al.*, 2020). Cependant, contrairement à d'autres représentations, un nuage de points ne contient pas d'informations sur la connectivité entre les points, ce qui conduit à des trous qui dégradent le photoréalisme du modèle 3D (Schütz *et al.*, 2020; Zerman *et al.*,

2020). Une structure avec de nombreux points est alors nécessaire pour obtenir un rendu photoréaliste, ce qui induit de grandes exigences en matière de calcul et de mémoire. Ces exigences sont particulièrement vraies pour les représentations de larges lieux en haute résolution. Étant donné que le traitement des nuages de points massifs requiert des ressources matérielles importantes, un utilisateur équipé d'un matériel moins performant comme un casque de réalité virtuelle autonome ne peut pas manipuler efficacement ces structures. Pour permettre à ces utilisateurs légers de visualiser des nuages de points massifs, l'idée de la visualisation à distance a été mise au point. L'idée est de stocker l'ensemble du nuage de points sur un serveur doté d'une grande capacité de traitement et que l'utilisateur client demande au serveur des informations sur les points à proximité de sa position. Nous proposons alors d'utiliser des images omnidirectionnelles avec carte de profondeur pour visualiser à distance des nuages de points statiques. Un client léger envoie une requête à un serveur pour obtenir un rendu omnidirectionnel d'un nuage de points distant à des coordonnées spécifiques sans recourir à une bande passante élevée et continue. L'image omnidirectionnelle retournée est directement visualisée en réalité virtuelle par l'utilisateur qui navigue sur le lieu d'un point de vue à l'autre par téléportation comme dans Google Street View (Anguelov *et al.*, 2010) ou QuickTime VR (Chen, 1995).

4.2.1 Approches Existantes

Pour visualiser à distance un nuage de points statique, vu à partir de la position de l'utilisateur, une représentation du nuage de points doit être transmise sur le réseau. Actuellement, deux types de représentations de nuages de points. La première approche, une approche géométrique 3D, consiste à transmettre les coordonnées des points individuellement avec leur propriété de couleur. La seconde, une approche de projection 2D, transmet une image correspondant à la projection des points.

L'idée de l'approche géométrique 3D est de transmettre un sous-ensemble de points visibles dans le champ de l'utilisateur afin d'éviter de transmettre l'ensemble du nuage de points. Le rendu du nuage de points est effectué localement côté utilisateur. Une manière efficace de représenter en mémoire le nuage de points est d'utiliser une organisation hiérarchique qui exploite la faible densité spatiale. L'espace 3D est décomposé en cubes englobants et les points sont encodés sous forme d'indices des cubes auxquels ils appartiennent. Cette stratégie a été standardisée par MPEG sous le nom G-PCC (*Geometry-based Point Cloud Compression*) (Cao *et al.*, 2021; Schwarz *et al.*, 2019) et recommande d'utiliser des structures comme des *k-d trees* ou des *octrees*. Cette approche de streaming de points est utilisée dans de nombreux systèmes comme Potree [1]

1. `https://potree.github.io`

ou GROOT (Lee *et al.*, 2020). Le principal inconvénient est que cette approche nécessite un matériel puissant côté client pour rendre des nuages de points volumineux. Une autre idée consiste à se passer d'un nuage de points dense en créant un maillage. Le passage à un maillage consiste à rajouter des informations de connectivité entre les points. Avec un maillage, les points d'une surface sont induits, ce qui évite au client de devoir manipuler des données trop volumineuses. Au lieu de transmettre un nuage de points dense, le serveur créé un maillage simplifié plus léger et plus facile à gérer sur le réseau. Cependant, la conversion d'un nuage de points massif en un maillage simplifié peut s'avérer difficile, en particulier en temps réel. Cette approche géométrique n'a pas été suivie car notre objectif était de pouvoir visualiser des nuages de points denses sur une architecture modeste.

D'autre part, l'idée de l'approche par projection 2D est de rendre le nuage de points du côté du serveur sur une image et de transmettre cette image à l'utilisateur. Chaque fois que le champ de vision de l'utilisateur est mis à jour côté client, le serveur envoie un nouveau rendu. Cette approche est censée être plus intéressante que l'approche géométrique car le nuage de points dense peut être représenté avec des données plus compactes. Mais la limitation principale est la nécessité d'une bande passante continue avec un débit élevé et une faible latence. En effet, un nuage de points dynamique, ou un changement de position ou d'orientation de l'utilisateur dans le lieu implique qu'un nouveau rendu doit être transmis. Une stratégie plus générale a été standardisée par MPEG sous le nom de V-PCC (*Video-based Point Cloud Compression*) (Cao *et al.*, 2021; Schwarz *et al.*, 2019), également adaptée à la transmission de nuages de points denses. L'idée est de projeter l'ensemble des points 3D sur un plan 2D du côté du serveur et de transmettre ce plan de projection à l'utilisateur pour que celui-ci puisse reconstruire localement le nuage de points. La plupart des approches V-PCC proposent de segmenter le nuage de points en patchs et de projeter ces patchs sur un atlas de texture transmis à l'utilisateur. Pour reconstruire la géométrie du nuage, ils proposent également de projeter des cartes de profondeur sur des patchs dans un atlas géométrique. La projection du nuage de points est alors similaire à l'utilisation de plusieurs caméras virtuelles enregistrant des parties du nuage de points et combinant ces images en une mosaïque (Graziosi *et al.*, 2020). Cependant, la plupart des méthodes V-PCC sont conçues pour transmettre des nuages de points correspondant à des objets, des éléments vus de l'extérieur. Ainsi, certaines propositions utilisent explicitement des caméras virtuelles placées à l'extérieur du nuage de points (Schwarz *et al.*, 2018) en configuration *outside-in* (sous-section 2.3.1), ce qui n'est pas directement utilisable pour visualiser un nuage de points d'un lieu vu de l'intérieur en configuration *inside-out*. Un avantage de l'approche par projection est que les techniques de traitement d'images traditionnelles sont applicables. La compression en image permet aussi une compression du nuage de points expérimentalement plus élevée

qu'avec l'approche géométrique (Zerman *et al.*, 2020). Ce taux de compression élevé pour les nuages de points denses s'explique par l'utilisation des techniques de compression d'images déjà bien connues en plus de la suppression naturelle des points occultés réduisant le nombre de points à transmettre (les points cachés ne sont pas projetés sur le rendu, seuls les points visibles sont transmis).

4.2.2 Approche Proposée

Notre proposition s'inscrit dans l'approche projection de l'état de l'art, à la différence que nous nous plaçons explicitement en configuration *inside-out* avec une caméra omnidirectionnelle à l'intérieur du nuage de points. L'image omnidirectionnelle permet à chaque visiteur de visualiser le nuage de points indépendamment de son orientation sans informations supplémentaires, ce qui réduit l'exigence d'une bande passante continue à haut débit pour un nuage de points statique. De plus, le rendu étant réalisé côté serveur, un nuage de points volumineux est aisément visualisé sur un matériel moins performant. Des interactions traditionnelles de la réalité virtuelle ont été développées, comme un mécanisme de navigation pour explorer le nuage de points à distance. Ce mécanisme permet aux utilisateurs de pointer avec un laser une zone du nuage de points et de s'y téléporter en envoyant une nouvelle requête au serveur. Un utilisateur navigue alors à travers le nuage de points par téléportations successives. Ce système propose des fonctionnalités supplémentaires par rapport au système du chapitre précédent comme la présence d'une parallaxe de mouvement ou l'ajout d'objets avec une gestion des occultations. Ces interactions dans la scène sont réalisées grâce à une carte de profondeur envoyée par le serveur en même temps que le rendu omnidirectionnel. L'ensemble du schéma de communication entre le serveur et le client est décrit figure 4.8.

4.2.3 Serveur

La vocation du serveur est de stocker un nuage de points volumineux, afin de décharger les visiteurs de la responsabilité de sa gestion en mémoire. Le serveur sélectionne alors les points à afficher et les convertis sous forme d'image omnidirectionnelle en fonction de la position d'un utilisateur. Cette sous-section détaille les algorithmes pour sélectionner les points avec la structure de données choisie et pour créer le rendu omnidirectionnel.

Octree

Le nuage de points est manipulé sur le serveur à l'aide d'un *nested octree* (Wimmer et Scheiblauer, 2006) pour affiner de manière adaptative le niveau de détail (*Level Of Details*, *LOD*) aux coordonnées d'un utilisateur. Un *nested octree* est une structure

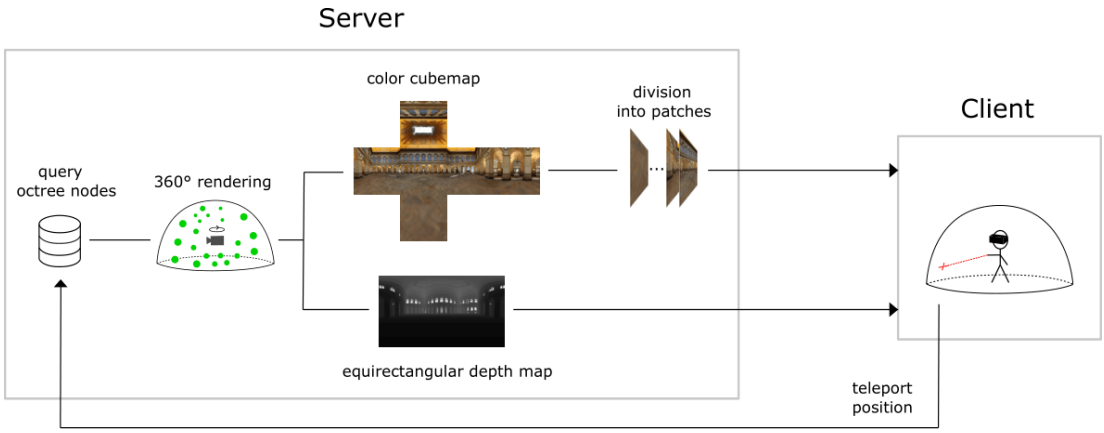


FIGURE 4.8 – Télé-immersion 3D 360° dans un nuage de points. Dans un premier temps, le serveur récupère les points visibles à la position de l'utilisateur. Les points sont ensuite projetés sur une image omnidirectionnelle avec une projection *cubemap* et une carte de profondeur omnidirectionnelle est créée. Les patchs de la *cubemap* et la carte de profondeur sont envoyés à l'utilisateur. L'utilisateur visualise l'image omnidirectionnelle en réalité virtuelle et sélectionne une destination de téléportation. Lorsque la commande de téléportation est déclenchée, une requête est envoyée au serveur pour récupérer le rendu du nuage de points aux coordonnées indiquées et le processus est répété.

de données arborescente où chaque nœud est attaché à des points réels du nuage. Il se distingue de l'*octree* classique où seules les feuilles de l'arbre sont réellement des points du nuage. Le serveur répond aux requêtes en envoyant une projection des points visibles aux coordonnées de l'utilisateur, considérablement plus légère en mémoire que le nuage de points complet. La principale différence avec l'approche V-PCC est que nous ne projetons pas les points du frustum de l'utilisateur sur un plan 2D, mais nous projetons tous les points visibles autour de l'utilisateur sur une sphère unité. Comme la transmission du rendu omnidirectionnel peut prendre du temps, le serveur envoie également un rendu temporaire inspiré des projections dépendantes de la fenêtre de visionnage (Chen *et al.*, 2018), afin de rendre rapidement une version dégradée du nuage de points.

Pour convertir le nuage de points en image omnidirectionnelle, la première étape est de parcourir les nœuds du *nested octree* pour récupérer les points pertinents étant donné la position de l'utilisateur. Un nœud du *nested octree* représente des points uniformément répartis dans une région donnée du nuage et les nœuds enfants représentent une partition cubique de cette région. Comme illustré figure 4.9 avec un *nested quadtreee* en 2D, les nœuds les plus profonds dans l'arborescence correspondent à un échantillonnage uniforme sur une partie de plus en plus restreinte de l'espace. Si l'utilisateur souhaite visualiser de plus près une partie du nuage de points, une augmentation du *LOD* est effectuée en allant en profondeur dans le *nested octree* pour les nœuds de cette région.

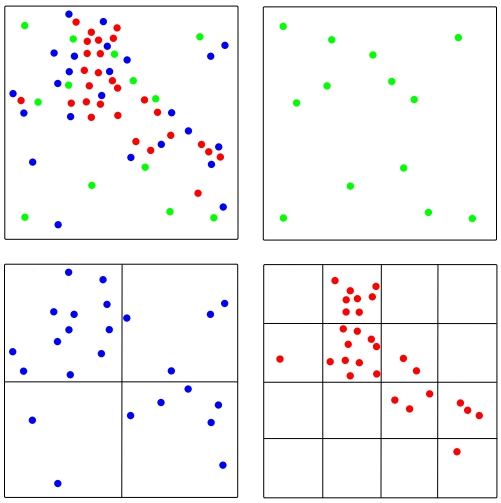


FIGURE 4.9 – Partition de points en 2D à l'aide d'un *nested quadtree*. L'ensemble des points Les points verts représentent le niveau de détail 0, les points bleus représentent le niveau de détail 1 et les points rouges représentent le niveau de détail 2. Haut-gauche : Ensemble de tous les points. Haut-droite : Points de la racine de l'arbre quadratique. Bas-gauche : Points dans les nœuds ayant une profondeur de 1 : Points dans les nœuds ayant une profondeur de 1. Bas-droite : Points dans les nœuds avec profondeur de 2.

Pour obtenir la relation hiérarchique, la structure nécessite un moyen d'évaluer la densité de points afin de déterminer si une région contient suffisamment de points pour créer un niveau de détail plus fin. Chaque région d'un nœud dans le nuage est alors divisée en une grille cubique de $c \times c \times c$, et un enfant du nœud est créé lorsque le nombre de points dans une cellule excède un certain seuil. Pour chaque nœud, un point représentatif est choisi pour correspondre au *LOD* du nœud. En choisissant le point le plus proche du centre de la cellule comme représentant, la distribution des points tend vers une distribution uniforme.

Image Omnidirectionnelle avec Profondeur

Après avoir sélectionné les points participants au rendu, le serveur doit les renvoyer sous forme d'image omnidirectionnelle avec carte de profondeur. Bien que la projection équirectangulaire soit la plus simple à obtenir, nous avons choisi d'utiliser une projection *cubemap* pour l'image couleur. Cette projection a l'avantage de se décomposer naturellement en patchs (sous-section 3.1.1). Au lieu de regrouper ces patchs sur la même image, ceux-ci sont envoyés indépendamment à l'utilisateur afin d'afficher progressivement un rendu partiel du nuage de points au lieu d'attendre l'image complète. D'autres projections en patchs peuvent être utilisées mais au prix d'un plus grand nombre de patchs, ce qui n'est pas désirable car augmentant le nombre d'images à envoyer sur le réseau. Lorsqu'un utilisateur demande au serveur le rendu du nuage de points, la requête contient

une position (x, y, z) ainsi qu'un angle d'orientation horizontale θ. À partir de ces informations, le serveur projette les points sur 6 patchs carrés à la position (x, y, z), la face avant dépendant de l'orientation θ. Le processus est similaire à la projection des points visibles sur six caméras perspectives. Un point n'est donc pas visible par l'utilisateur s'il n'appartient à aucun champs de vision des six caméras.

La première étape de la création d'un patch consiste à sélectionner les points à projeter sur la caméra. Cette sélection est effectuée par un parcours en profondeur sur le *nested octree* avec une file de priorité. Au départ, la file ne contient que le point à la racine du *nested octree*. À chaque itération, le nœud ayant la plus grande visibilité est sélectionné dans la file de priorité afin d'afficher d'abord les plus grands nœuds proches de la caméra. Le facteur de visibilité v d'un nœud à la profondeur n dans l'*octree*, dépendant du rayon du nœud r_n et de la distance d entre le nœud et la caméra est donné par la formule :

$$\begin{aligned} v &= \frac{r_n}{d} \\ r_n &= \frac{r}{2^n} \end{aligned} \tag{4.2}$$

Le rayon r_n est calculé comme une fraction d'un rayon global r contenant tous les points du nuage de points à partir de l'origine du *nested octree*. Lorsqu'un nœud est sélectionné, les points qu'il contient sont ajoutés dans une liste des points à rendre. Les nœuds enfants sont ajoutés à la file de priorité si la taille de projection des points qu'ils contiennent n'est pas inférieure à un pixel. La taille de projection p en pixels d'un point sur un patch est donnée par l'équation suivante, dérivée de la formule générale de projection perspective d'un point (Kang *et al.*, 2019) :

$$p = \frac{s\, r_n}{2\, c\, d} \tag{4.3}$$

où s est la taille du patch carré. Un rayon de point de $\frac{r_n}{c}$ est utilisé au lieu de r_n pour estimer la taille du pixel car un nœud du *nested octree* contient les points sur une grille de taille c. Ce processus de sélection est répété jusqu'à ce qu'une condition d'arrêt soit atteinte. Dans les visualiseurs de nuages de points traditionnels, l'algorithme s'arrête lorsque la liste des points à afficher atteint un budget. Un budget de points limité maintient un taux de rafraîchissement interactif particulièrement important en réalité virtuelle. Le *LOD* est alors déterminé par le budget de points. Dans notre cas, un budget de points beaucoup plus élevé est utilisé car nous n'avons pas de contraintes d'interactivité sur le serveur. Le parcours en profondeur du *nested octree* se poursuit jusqu'à ce qu'un nœud avec une taille de projection inférieure à un pixel soit atteint. À ce moment, l'algorithme s'arrête car les points de ces nœuds et de leurs enfants ne sont plus visibles sur une image. La taille d'un patch de la *cubemap* donne alors un niveau

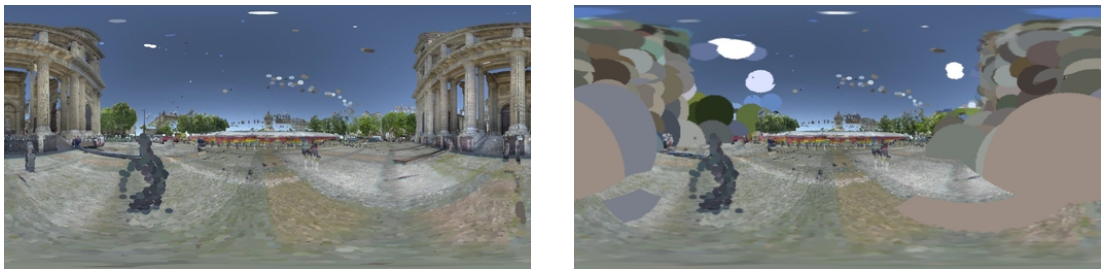


FIGURE 4.10 – Rendu omnidirectionnel d'un nuage de points. Gauche : Rendu définitif. Droite : Rendu temporaire.

de profondeur maximum à ne pas dépasser dans le *nested octree*, le *LOD* est déterminé par cette résolution. Ce critère de raffinement est un moyen efficace de sélectionner tous les points qui contribuent effectivement au rendu.

L'étape finale consiste à projeter sur une image les points sélectionnés. Pour éviter les images éparses, un algorithme de rendu en *splats* est utilisé. Cette méthode consiste à étaler la taille d'un point sur une surface de plusieurs pixels pour couvrir les espaces entre les points. Comme les points sont uniformément distribués, une taille de *splat* constante est utilisée, donnée par l'équation 4.3. Pour créer une image, les *splats* sont traités dans un *shader* de rastérisation comme les facettes d'un maillage, également en charge du traitement des occultations entre les points. Les processus de sélection et de rendu sont répétés six fois pour obtenir tous les patchs de la *cubemap*. Un patch individuel est directement envoyé à l'utilisateur lorsque son rendu est terminé. La figure 4.10 à gauche donne un exemple d'image omnidirectionnelle obtenue avec cette approche.

Pour compléter l'image couleur omnidirectionnelle, la carte de profondeur est aussi générée. Cette carte de profondeur omnidirectionnelle est obtenue grâce à une passe de rastérisation pour connaitre la distance entre les *splats* et la caméra. Après génération, la carte de profondeur est transmise au client, mais contrairement à l'image couleur elle n'est pas divisée en patchs. Pour la transmission, la carte de profondeur est envoyée en une fois au client avec une projection équirectangulaire. Nous avons choisi d'envoyer la carte de profondeur en une seule image, car la manipulation d'une carte de profondeur partielle n'est pas intéressante pour les interactions. Du côté client, l'utilisateur attend de recevoir la carte de profondeur pour interagir avec la scène.

Rendu Temporaire

La transmission des patchs d'une image omnidirectionnelle de haute qualité avec la carte de profondeur associée peut prendre du temps avec une faible bande passante. Ce temps de latence pour l'utilisateur pendant la téléportation altère sa qualité d'expérience et son sentiment de présence. Pour éviter ce problème, l'idée est d'envoyer

rapidement un rendu temporaire afin d'avoir un retour visuel au plus tôt après la demande de téléportation. Le serveur envoie alors un premier rendu temporaire au client, affiché en attendant de recevoir le rendu définitif. L'ordre d'envoi des données est le suivant : le serveur envoie le rendu temporaire, puis la carte de profondeur, et enfin le rendu définitif. L'utilisateur peut interagir avec la scène, notamment se déplacer, juste après avoir reçu la carte de profondeur. Le rendu temporaire est inspiré des projections dépendantes de la fenêtre de visionnage des images omnidirectionnelles qui réduisent la qualité l'image omnidirectionnelles aux régions hors du champ de vision de l'utilisateur (sous-section 3.1.1). Notre rendu temporaire correspond à une projection *cubemap* avec des patchs ayant une résolution réduite par rapport au rendu définitif (ressemblant à l'idée de la projection *offset-cubemap*). Cette réduction de résolution induit automatiquement une réduction du *LOD* étant donné notre critère de raffinement. Le parcours du *nested octree* n'étant alors pas aussi profond qu'avec le rendu définitif, le temps pour générer le rendu d'un patch est donc réduit. Étant donné qu'il est souhaitable de maintenir une qualité élevée dans la région vers laquelle le regard de l'utilisateur est dirigé, une haute résolution est conservée pour la face avant, la résolution est dégradée seulement sur les autres patchs. Le patch avant n'est donc pas envoyé pour le rendu final. Du point de vue de l'utilisateur, le rendu temporaire est une image alors omnidirectionnelle dont la région frontale en haute qualité et les autres régions en basse qualité. Les patchs de faible qualité sont progressivement remplacés par des patchs de haute qualité quand ils sont reçus côté client. La figure 4.10 à droite donne un exemple de rendu temporaire.

4.2.4 Client

De nombreuses métaphores de navigation ont été développées dans le cadre de la réalité virtuelle pour se déplacer dans une scène (Bowman *et al.*, 2001). Cependant, pour naviguer dans une scène distante, la métaphore à utiliser dépend directement du schéma de communication entre le serveur et le client. Dans une application de streaming (streaming de points ou streaming de rendu), le serveur rend continuellement l'environnement aux coordonnées de l'utilisateur. Ainsi, une métaphore de déplacement par direction qui permet à l'utilisateur de spécifier en continu la direction de son mouvement semble donc appropriée. En revanche, lorsque le serveur envoie ponctuellement le rendu du point de vue de l'utilisateur, comme notre serveur, une métaphore de déplacement basée cible est mieux adaptée à cette communication discrète. En effet, ce type de métaphore ne communique que la destination finale plutôt que les détails de la trajectoire complète, évitant alors la transmission continue des informations sur la trajectoire. Une métaphore de téléportation locale a donc été développée pour que le client puisse naviguer dans le nuage de points distant de manière cohérente avec notre schéma de communication du serveur. Du côté des utilisateurs, les patchs de la *cubemap* du

FIGURE 4.11 – Vue d'un utilisateur en réalité virtuelle du système de télé-immersion 3D 360°. Le nuage est une fusion de scans avec 730 millions de points. Le rendu est effectué du côté du serveur et est envoyé sous la forme d'image omnidirectionnelle. Les patchs *cubemap* de l'image omnidirectionnelle ont une résolution de 1024×1024 pixels. L'utilisateur pointe un laser au sol pour sélectionner la destination à laquelle il souhaite se téléporter.

nuage de points sont directement affichés dans l'ordre où ils sont reçus. Après réception de la carte de profondeur, un visiteur lance un rayon laser vers une destination, comme la métaphore traditionnelle de la téléportation (figure 4.11). La carte de profondeur est d'abord utilisée pour vérifier si la surface est plate aux coordonnées du pixel pointé avec un calcul de la normale, puis utilisée pour calculer les coordonnées globales de ce pixel dans le nuage de points. Lorsque la commande de téléportation est déclenchée, une requête est envoyée au serveur pour récupérer le rendu à ces coordonnées globales avec un angle θ correspondant à l'orientation horizontale de l'utilisateur. La figure 4.11 montre exemple d'utilisation de la téléportation dans le nuage de points.

La carte de profondeur avec l'image omnidirectionnelle est aussi utilisée afin d'améliorer l'immersion et l'interaction de la scène côté client comme présenté sous-section 4.1.3. Après avoir reçu le rendu final, la carte de profondeur omnidirectionnelle est utilisée pour reconstruire partiellement le lieu en 3D vu de la position de l'utilisateur. Le rendu avec occultations permet d'avoir des incarnations des autres visiteurs pouvant être occultées par le lieu ou d'ajouter des objets virtuels (figure 4.3). La parallaxe de mouvement aussi est intéressante car elle permet de se déplacer dans la scène uniquement à partir de la carte de profondeur (figure 4.4). L'avantage de cette parallaxe est alors qu'elle ne nécessite pas de nouvelle requête au serveur contrairement aux approches en streaming. En revanche, même si en théorie la taille de la parallaxe n'est pas bornée, la dégradation de la qualité du rendu quand on s'éloigne de la position de la caméra peut inciter les utilisateurs à limiter leurs déplacements. Aussi, afin que l'utilisateur soit immergé dans le scan avec les dimensions correctes, un facteur d'échelle doit être appliqué à la reconstruction (équation 4.1). Comme développé plus haut, les points du nuage à la

position de l'utilisateur sont capturés à l'aide de caméras perspectives paramétrées par une distance maximale définissant le champ de vision des points visibles. Cette distance maximale, la même pour les six caméras, est alors utilisée comme valeur d'échelle.

4.2.5 Évaluation

Notre système est conçu pour que des clients légers en réalité virtuelle puissent être télé-immergés sur un nuage de points volumineux. Un point critique est donc de garantir que la bande passante est utilisée avec modération et que la quantité de données à transmettre est relativement faible. Pour vérifier cela, nous avons conçu une expérience qui compare notre approche avec d'autres approches standards.

Protocole

Pour évaluer l'utilisation de la bande passante, nous avons mesuré son évolution en comparant notre approche avec le streaming de points (approche géométrique) et le streaming de rendu (approche projection) présentés sous-section 4.2.1. Nous avons choisi de comparer les approches en fonction de la variation de la bande passante et de la quantité totale de données envoyées. Pour ce faire, nous avons mesuré la bande passante utilisée par chaque méthode au cours d'une session d'environ 1 minute. Une session consiste à téléporter un client à 4 positions prédéfinies et à effectuer 2 rotations horizontales pour chaque position, avec une pause de 5 secondes entre chaque changement. Le test a été réalisé sur un nuage à 730 millions de points correspond à une fusion de scans LiDAR du Palais Brongniart à Paris. Nous avons utilisé des patchs carrés sur 32 bits pour l'image omnidirectionnelle, avec des tailles de 512 pixels pour la basse qualité et de 1024 pixels pour la haute qualité. Nous avons utilisé une carte de profondeur sur 16 bits avec une largeur de 3840 pixels et une hauteur de 2160 pixels. Les patchs et la carte de profondeur sont envoyés au format JPG. Les serveurs et clients ont été développés avec Unity. Le streaming de points a été mis en œuvre à l'aide d'un serveur avec un *octree* sur WebRTC. Les nœuds à l'intérieur du champ de vision du client sont requêtés au serveur, le serveur répond en envoyant les coordonnées et les couleurs des points compressés avec Draco [2]. Le serveur de streaming de rendu a été développé avec l'application Render Streaming de Unity, également basée sur WebRTC. Le rendu du point est effectué côté serveur et les images sont directement transmises au client, compressées avec le codec h264. Nous avons utilisé les dimensions par défaut de la vidéo diffusée, à savoir une largeur de 1280 pixels et une hauteur de 720 pixels. La bande passante pour les trois approches a été mesurée avec le logiciel d'analyse de réseau Wireshark en capturant uniquement les données reçues du serveur de nuages de points.

2. `https://github.com/google/draco`

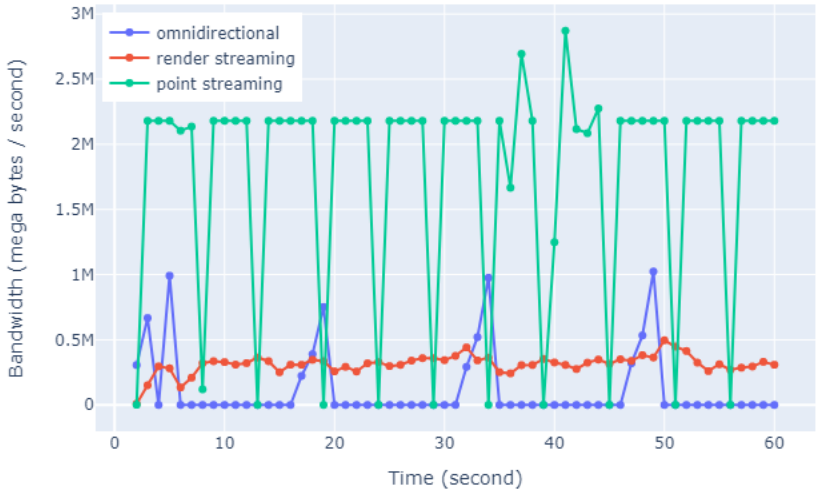


FIGURE 4.12 – Utilisation de la bande passante des approches.

Résultats

Les valeurs de la bande passante mesurée sont présentées figure 4.12. On observe que la bande passante est bien utilisée en continu pour les deux approches de streaming tandis qu'avec notre approche elle est utilisée ponctuellement. Le serveur de streaming de points doit envoyer de nouveaux points pour affiner le rendu après un changement de position et d'orientation, tandis que le serveur de streaming de rendu doit envoyer l'image vidéo actuelle à une fréquence élevée. Pour notre approche, le graphique montre une utilisation ponctuelle de la bande passante avec des pics correspondant aux changements de position, sans données supplémentaires reçues lors du traitement des rotations comme anticipé. En additionnant toutes les quantités de données, nous constatons que le serveur de streaming de points envoie environ 15 fois plus d'octets que notre approche, et que le serveur de streaming de rendu envoie environ 2.5 fois d'octets que notre approche. Ce résultat montre que notre méthode utilise également moins de données, même en prenant en compte les données supplémentaires (patchs du rendu temporaire et carte de profondeur). Notre méthode est alors bien adaptée pour une télé-immersion légère dans un nuage de points statique. Cependant, contrairement à notre approche, les approches par streaming peuvent être utilisées pour visualiser des nuages de points dynamiques.

4.3 Conclusion

Dans ce chapitre, nous avons exploré l'apport de la carte de profondeur pour une télé-immersion 3D 360°. L'addition de l'information 3D à l'image omnidirectionnelle permet de passer d'une représentation d'éléments en champ lointain à une représentation d'éléments en champ proche. Avec la carte de profondeur, notre système est capable de télé-immerger des utilisateurs, non plus simplement sur des lieux extérieurs mais dans des lieux intérieurs. En exploitant celle-ci sous forme de proxy, un volume sur lequel on projette l'image omnidirectionnelle, de nouvelles fonctionnalités en termes d'immersion et d'interaction sont possibles. La gestion des occultations, la sensation que les objets 3D ne flottent pas dans l'environnement, et surtout la possibilité de naviguer librement, constitue un progrès notable par rapport au chapitre précédent. Cette amélioration a été mise à profit pour le développement d'un système de télé-immersion sur des scans 3D. Ce système permet à plusieurs visiteurs de se retrouver sur un nuage de points volumineux statique. Grâce à une utilisation réduite de la bande passante, ce système offre une bonne solution de télé-immersion pour réunir des étudiants sur un lieu d'intérêt comme une salle de classe.

Néanmoins, comme relevé, nous baser sur des images omnidirectionnelles avec des cartes de profondeur n'est pas suffisant pour répondre aux besoins de notre système de télé-immersion 3D. Tout d'abord, nous n'avons actuellement pas de méthode d'estimation de carte de profondeur en temps réel donnant des reconstructions avec une estimation fine de la géométrie sans artéfacts. Les reconstructions actuelles perdent le relief de certains éléments du lieu ou ont des distorsions qui rendent des surfaces plates ondulées. Avec une unique caméra omnidirectionnelle statique, une estimation de la géométrie pour chaque image d'un flux en direct n'est aujourd'hui pas envisageable et donc un système de télé-immersion 3D 360° sur des lieux dynamiques n'existe pas avec cette approche. De plus, l'image omnidirectionnelle, même avec la carte de profondeur, reste par essence une représentation associée à l'environnement de la scène. L'image omnidirectionnelle définit simplement les limites de la scène et il n'est pas possible d'interagir individuellement avec un élément de l'image (comme le sélectionner ou se déplacer autour). L'image omnidirectionnelle contient seulement des éléments d'arrière-plan et une manière de modéliser des objets (les éléments de premier plan) doit être trouvée. Le prochain chapitre se penche sur la résolution de ces problèmes et proposera une nouvelle représentation 3D de la scène, supportant une modélisation des objets et pouvant être obtenues en temps réel.

Chapitre 5

Télé-Immersion 3D 360° Dynamique

Dans le chapitre précédent, nous avons exploré les possibilités offertes par les cartes de profondeur pour mettre au point des systèmes de télé-immersion dans des lieux simples. Mais cette représentation est trop simple pour modéliser de manière satisfaisante les lieux plus complexes mélangeant environnement et objets. De plus, nous nous sommes concentrés sur la télé-immersion sur des lieux statiques, la contrainte visant à obtenir cette représentation en temps réel n'a pas été considérée. Dans ce chapitre, nous proposons une nouvelle représentation 3D pour modéliser les lieux comportant environnements et objets, et comment la générer en temps réel. Cette représentation est la base de notre dernière itération du système de télé-immersion développé. En utilisant cette version, un enseignant peut tenir des cours en ligne simplement grâce à une caméra 360° avec des étudiants plongés dans la salle de classe en réalité virtuelle depuis chez eux. Avec la représentation 3D en temps réel, les étudiants se déplacent librement et interagissent avec l'enseignant et les autres étudiants sur place. Ce dernier système de télé-immersion permet de réaliser un scenario où l'enseignant donne un cours sur un objet d'intérêt avec des étudiants distants qui examinent cet objet sous divers angles. Afin que le système soit facilement déployable, nous avons mis l'accent sur une utilisation réduite de la bande passante du réseau, comparable à celle d'un logiciel de visioconférence.

5.1 Représentations 3D 360°

Notre objectif initial est de mettre au point un système de télé-immersion asymétrique où des hôtes et des visiteurs peuvent être rassemblés sur un lieu commun capturé uniquement avec une caméra omnidirectionnelle statique. L'idée est qu'avec une caméra omnidirectionnelle posée dans le lieu, les visiteurs sont immergés en réalité virtuelle et peuvent se déplacer librement. Les hôtes voient les visiteurs en réalité mixte sur le lieu, incarnés en avatar 3D à leurs positions dans la scène virtuelle. Un tel système serait une

contribution importante pour l'enseignement en ligne, des cours pourraient se tenir pour étudier un objet particulier, comme un moteur ou une machine en cours d'ingénierie, et les élèves pouvant se déplacer autour de l'objet ou sélectionner des parties pour avoir des explications sur son fonctionnement.

Comme précédemment, le principal problème pour atteindre ce type de système réside dans la création d'une parallaxe à partir de la vue omnidirectionnelle afin que les utilisateurs puissent s'éloigner de la position de la caméra. Au chapitre 3, nous avons argumenté que l'utilisation d'une image omnidirectionnelle n'est pas suffisante car des informations géométriques sont nécessaires pour une navigation libre sur le lieu. Au chapitre 4, nous avons ajouté à l'image omnidirectionnelle une carte de profondeur pour modéliser la géométrie du lieu. Nous avons conclu que cette représentation offrait bien la possibilité de se déplacer librement mais se révélait moins pertinente lorsque le lieu est occupé par des éléments de premier plan. En effet, la carte de profondeur est en essence une unique grande surface 3D et les éléments 3D indépendants ne peuvent pas être modélisés. La question est alors de trouver une représentation qui supporte une parallaxe de mouvement pour des lieux relativement complexes et non plus pour des lieux simples comme une salle vide. Dans cette section, nous allons présenter les représentations existantes pour des lieux complexes et notre proposition.

5.1.1 Approches Existantes

Dans une scène créée exclusivement à partir d'une image omnidirectionnelle, les possibilités de navigation sont fortement limitées. Afin qu'un utilisateur de notre système puisse s'éloigner de la position de la caméra, il faut pouvoir créer une parallaxe suffisamment importante de manière à ce qu'il puisse de se déplacer librement. Dans le chapitre précédent, nous avons obtenu cette parallaxe à l'aide de cartes de profondeur estimé directement sur l'image omnidirectionnelle. L'association d'une image omnidirectionnelle avec une carte de profondeur permet d'obtenir un maillage 3D du lieu assurant une parallaxe sur l'ensemble de la scène. Mais cette représentation relativement simple présente des inconvénients. Cette représentation ne peut pas être utilisée pour la téléimmersion dans des lieux dynamiques car les méthodes ne sont pas temps réel et elle ne permet pas de modéliser une scène avec des objets autour desquels il serait possible de se déplacer. Ceci est la conséquence du fait qu'elle permet uniquement de modéliser un environnement. La capacité de la carte de profondeur à ne modéliser que des scènes simples a alors pousser au développement de nouvelles représentations 3D 360° plus élaborées obtenues à partir d'une vue omnidirectionnelle.

Une alternative pour modéliser des scènes avec une topologie plus complexe consiste à recourir aux représentations stratifiées. Une représentation stratifiée est un ensemble

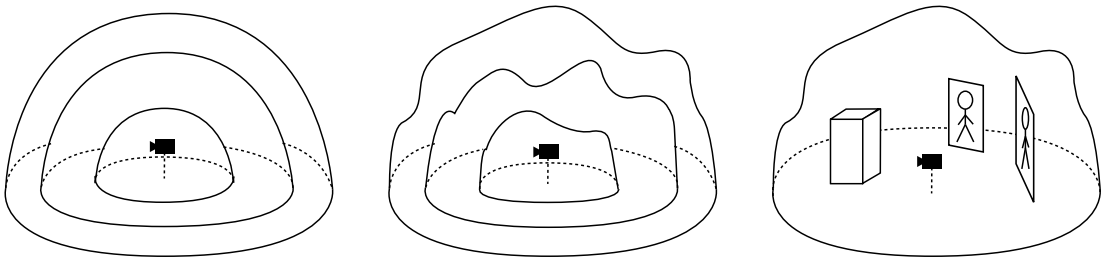


FIGURE 5.1 – Représentations 3D 360°. Gauche : Représentation multi-sphères. Milieu : Représentation multi-sphères avec cartes profondeur. Droite : Notre représentation avec carte de profondeur omnidirectionnelle, modèle 3D et *billboard*.

d'images semi-transparentes du lieu prises depuis le même point de vue, avec ou sans modélisation explicite de la géométrie (Richardt *et al.*, 2020). Les représentations stratifiées sont vues comme en 2.75D car entre la carte de profondeur en 2.5D et la reconstruction 3D complète (Dhamo *et al.*, 2019). Étant donné qu'une représentation stratifiée contient des informations sur les éléments dissimulés derrière les occultations, elle est capable de modéliser des topologies plus complexes qu'un convexe étoilé. Lorsque la géométrie est implicite, sans cartes de profondeur, l'ordre des couches contient l'information géométrique. Si une couche se trouve devant une autre, cela indique que cette couche est plus proche de la caméra. Dans ce cas à géométrie implicite, la représentation est appelée image multi-plans (MPI) (Han *et al.*, 2022; Tucker et Snavely, 2020). Initialement conçue pour les images perspectives traditionnelles, des auteurs ont adapté la représentation MPI pour fonctionner avec les images omnidirectionnelles en introduisant la représentation en image multi-sphères (MSI) (Attal *et al.*, 2020; Broxton *et al.*, 2020). La représentation MSI consiste en un ensemble de couches sphériques concentriques semi-transparentes centrées sur la position de la caméra (figure 5.1 gauche). L'utilisation de plusieurs couches génère un effet de parallaxe en réponse aux mouvements de l'utilisateur, les couches les plus proches de la caméra se déplaçant sur une plus grande distance que les couches plus éloignées. L'idée est similaire aux mosaïques concentriques (Shum et Kang, 2000) qui sont capturées grâce à des caméras à rotation autour d'un axe décentrée avec des rayons différents Pour générer la MSI, (Broxton *et al.*, 2020) s'appuient sur un dispositif multi-caméras spécialement conçu tandis que (Attal *et al.*, 2020) se basent sur une simple paire d'images stéréo omnidirectionnelle ODS avec un réseau de neurones inférant en temps réel les différentes couches. De manière similaire, (Waidhofer *et al.*, 2022) proposent le concept d'image multi-cylindres (MCI), un ensemble de couches cylindriques calculées à partir d'une seule image panoramique. Cependant, cette représentation requiert un nombre de couches important afin que la structure en couches ne devienne pas apparente sous forme d'artéfacts lorsque l'utilisateur s'éloigne du centre. Une stratégie pour réduire les artéfacts lors de mouvements importants est

alors de modéliser explicitement la géométrie à l'aide d'une carte de profondeur par couche. Cette stratégie conduit à la représentation en images de profondeur stratifiées (LDI) (Kopf *et al.*, 2020; Shih *et al.*, 2020; Shade *et al.*, 1998). À la différence des MPI, l'utilisation de différentes couches ne représente pas implicitement des données de profondeur mais plutôt des informations topologiques : si deux éléments se situent sur des couches différentes, ils sont considérés comme indépendants. Cette représentation s'étend alors aux images omnidirectionnelles en utilisant des images multi-sphères avec une carte de profondeur omnidirectionnelle pour chaque couche (figure 5.1 milieu). Une approche courante pour les images omnidirectionnelles est d'employer une LDI sphérique avec seulement deux couches, afin de représenter les éléments de premier plan et d'arrière-plan (Hedman *et al.*, 2017). Dans (Mühlhausen *et al.*, 2020; Serrano *et al.*, 2019), une LDI sphérique en trois couches est proposée : une couche de premier plan dynamique, une couche extrapolée pour les régions d'arrière-plan statiques occultées par les objets en mouvement et une couche de remplissage pour combler les régions occultées par les objets statiques. (Lin *et al.*, 2020) ont proposé le panorama multi-profondeurs (MDP), une LDI panoramique avec un nombre arbitraire de couches. Le MDP est obtenu à partir des images d'une caméra multi-objectifs passées en entrée d'un réseau de neurones prédisant les couches d'une MPI. Malheureusement, la majorité des représentations stratifiées cherchent à étendre la capacité de déplacement de quelques centimètres afin d'obtenir une parallaxe de mouvement de la tête pour un visionnage plus immersif des vidéos omnidirectionnelle, une représentation *3-DoF+* (Jung *et al.*, 2018). Une autre représentation est donc à trouver pour une navigation *6-DoF* complètement libre à travers la scène. La représentation stratifiée semble aussi déconseillée pour un système de télé-immersion avec une bande passante raisonnable étant donné la quantité de données supplémentaires à transmettre pour chaque image du flux vidéo.

Les progrès récents dans le domaine du rendu neuronal, c'est-à-dire l'utilisation de modèles d'apprentissage profonds pour la génération d'images réalistes, ont également ouvert de nouvelles perspectives pour la génération de parallaxe. Une avancée majeure a été la proposition du réseau de neurones NeRF (*Neural Radiance Fields*) (Mildenhall *et al.*, 2021) qui permet de représenter une scène sous forme de fonction, prenant en entrée une position et une orientation, et produisant en sortie un rendu de la scène. En optimisant cette fonction, un NeRF est capable de produire de nouvelles vues photo-réalistes de scènes à l'apparence et à la géométrie complexes. Initialement, un NeRF est entraîné à partir d'un ensemble d'images perspectives de la scène, mais des auteurs ont proposé des architectures pour générer des nouvelles vues d'une scène à partir d'une unique image omnidirectionnelle. (Hara et Harada, 2022; Kulkarni *et al.*, 2022; Hsu *et al.*, 2021) ont notamment proposé de générer de nouvelles vues avec un NeRF à partir d'une image omnidirectionnelle avec la carte de profondeur associée. Étant donné

qu'une image et sa carte de profondeur permettent de créer un nuage de points avec un rendu épars (sous-section 4.1.2), l'idée est de recourir à un NeRF pour remplir les zones vides et obtenir un rendu dense. Ce principe de complétion du nuage de points par rendu neuronale a aussi été exploré en utilisant d'autres réseaux que NeRF en ajoutant d'autres sources d'informations comme l'agencement du lieu ou la sémantique (Koh *et al.*, 2021; Xu *et al.*, 2021a). Toutefois, même si ces approches offrent théoriquement une plus grande parallaxe, les approches de rendu neuronal ne sont pas encore adaptées pour la télé-immersion, la majorité actuellement ne supportant pas un rendu en temps réel et ne se généralisant pas aux scènes dynamiques. De plus, ces approches ont tendance à halluciner, c'est-à-dire à ajouter au rendu des éléments qui n'existent pas réellement, ce qui est dommageable pour un système de télé-immersion où la scène doit être une copie fidèle du lieu.

5.1.2 Approche Proposée

Il est commun pour une scène de réalité virtuelle de considérer deux types d'objets : les objets d'intérêt et les objets ambiants (Lee *et al.*, 2021). Les objets d'intérêt sont les éléments avec lesquels l'utilisateur va interagir tandis que les objets ambiants sont des éléments de décor qui peuplent l'environnement sans réelle fonction. Pour notre application d'enseignement à distance, nous utilisons cette distinction et proposons qu'une session se concentre sur l'étude d'un objet d'intérêt, les utilisateurs sur site ne seront alors que des objets ambiants. L'objet central de la session devra avoir une représentation 3D volumétrique afin qu'un utilisateur puisse interagir correctement avec, mais on permettra que la représentation des hôtes puisse être dégradée car ils ne seront pas les éléments principaux.

Afin de télé-immerger des visiteurs dans un lieu, à partir d'un flux vidéo omnidirectionnelle en direct, en leur permettant de se déplacer librement, nous nous sommes inspirés des représentations stratifiées. Nous adoptons une variante de la représentation stratifiée à deux couches qui modélisent le premier plan et l'arrière-plan. Une piste intéressante est celle explorée dans Tour Into Picture (Horry *et al.*, 1997), un algorithme qui permet la création d'un modèle 3D à partir d'une seule image perspective. Dans cet algorithme, des représentations 3D distinctes sont utilisées pour le premier plan et l'arrière-plan. L'arrière-plan est modélisé à l'aide d'un maillage, estimé grâce à des indices visuels sur la perspective (similaire aux méthodes d'estimation d'agencement sous-section 4.1.4), tandis que les éléments de premier-plan sont rendus sous forme de *billboards*, c'est-à-dire des plans 2D texturés positionnés dans l'espace 3D. Cette représentation en *billboard* est appropriée lorsqu'un objet 3D est créé à partir d'une seule vue, mais elle ne permet pas de générer de nouveaux points de vue autour de celui-ci en raison du manque d'information. Une autre inspiration a été la proposition de

(Dupont de Dinechin et Paljic, 2019) qui traite aussi différemment l'arrière-plan et les personnes pour créer des avatars 3D dans un environnement 3D. Cependant, la méthode ne permet pas d'obtenir ces avatars en temps réel. Nous avons alors suivi la représentation de Tour Into Picture, en représentant l'environnement avec un maillage et les hôtes avec des *billboards*. Étant donné que le *billboard* n'est pas adéquat pour modéliser les objets d'intérêt de la scène (on ne peut pas avoir de nouvelles vues du *billboard* en se déplaçant autour) nous proposons d'utiliser des modèles 3D pour les représenter. Cette représentation au complet est illustrée figure 5.1 à droite. L'avantage est que les objets d'intérêt sous forme de modèle 3D peuvent être traités comme des objets classiques de réalité virtuelle avec les mêmes interactions (sélection et manipulation). À noter que le cas d'usage visant à étudier un objet précis, il est possible que le même objet soit étudié sur plusieurs sessions (comme un cours portant sur le même objet au fil des années). Il parait pertinent dans ce cas qu'un modèle 3D de cet objet puisse être déjà connu à l'avance. Il n'est pas donc nécessaire de procéder à chaque fois à la reconstruction 3D volumétrique de l'objet d'intérêt pour obtenir notre représentation. Nous proposons de l'ajouter à notre représentation avec un recalage de son modèle 3D connu sur la vidéo omnidirectionnelle.

5.2 Télé-Immersion 3D 360° Temps Réel

Pour la version définitive de notre système de télé-immersion, nous nous basons sur cette nouvelle représentation 3D 360°. Dans notre cas d'usage, nous considérons la salle de classe comme un lieu dynamique. Un visiteur doit pouvoir interagir avec l'objet d'intérêt et communiquer directement avec les hôtes, l'obtention de la représentation 3D 360° en temps réel est alors essentielle. Cette section présente le fonctionnement de cette dernière version, notamment comment les différents éléments qui composent la scène sont reconstruits. La méthode de reconstruction sera ensuite évaluée en termes de performance et de qualité visuelle. Enfin, les limites de notre système seront évoquées avec des pistes d'amélioration.

5.2.1 Fonctionnement

Avec la représentation 3D 360° proposée, nos besoins pour notre système de télé-immersion sont remplis : l'utilisateur se déplace librement dans la scène et il peut interagir avec l'objet d'intérêt. Cependant, pour se télé-immerger sur un lieu dynamique, cette représentation doit être reconstruite en temps réel pour que l'utilisateur ait une scène qui corresponde à l'état du lieu sans décalage de temps important. Comme la génération de cette représentation peut être coûteuse en temps, nous avons suivi une stratégie classique qui consiste à traiter séparément les éléments statiques et dynamiques

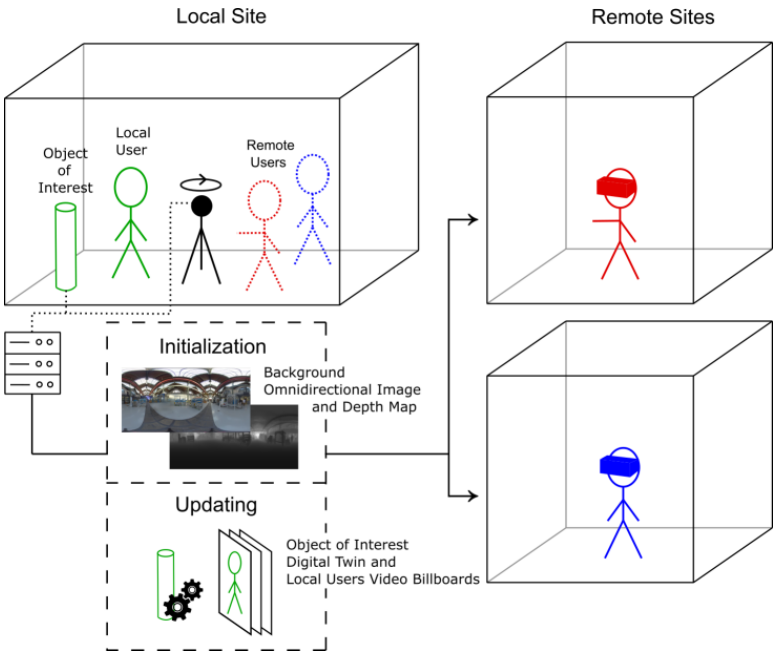


FIGURE 5.2 – Télé-immersion 3D 360° temps réel. Les visiteurs se connectent au serveur pour naviguer librement dans la vidéo omnidirectionnelle diffusée. Un visiteur reçoit à l'initialisation les éléments statiques du lieu, et reçoit au fur et à mesure la mise-à-jour des éléments dynamiques.

dans l'image dans le but d'obtenir un algorithme en temps réel.

Conception

Pour concevoir un algorithme temps réel, nous avons séparer notre processus de reconstruction 3D en deux étapes, une phase d'initialisation pour reconstruire les éléments statiques et une phase d'actualisation pour la mise à jour des éléments dynamiques. Le système de télé-immersion proposé avec cette approche de reconstruction est présenté figure 5.2.

Pour simplifier la reconstruction, nous faisons l'hypothèse que l'environnement est statique et que seuls les objets sont dynamiques (l'intérêt d'un environnement dynamique étant limité). Ainsi, la reconstruction de l'environnement peut être réalisée en une seule fois à l'initialisation. Nous supposons aussi que l'objet d'intérêt a une position fixe dans l'espace et qu'il ne se déplace que faiblement. Ceci est pertinent car l'objet d'intérêt est une machine ou mécanisme encombrant dans notre cas d'usage. Cette supposition permet de retrouver sa localisation globale seulement à l'initialisation. Si sa position change, elle peut être modifiée relativement à sa position d'origine dans l'étape d'actualisation. La création de la scène débute alors par la reconstruction de l'environnement en 3D et le recalage de l'objet d'intérêt. Cette étape d'initialisation est résumée

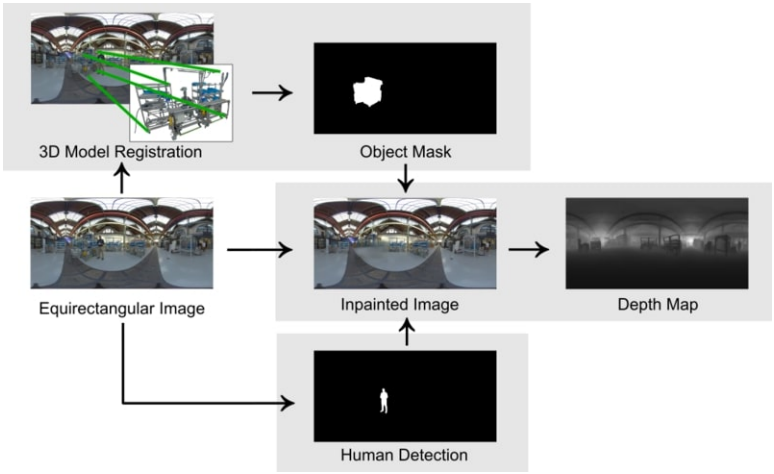


FIGURE 5.3 – Initialisation de la reconstruction 3D 360°. Les informations sur l'objet d'intérêt (boîte du haut) et sur les personnes (boîte du bas) sont nécessaires pour éliminer le premier plan de l'environnement (boîte du milieu).

figure 5.3.

Pour obtenir une reconstruction 3D complète de la scène, les éléments dynamiques doivent être ajoutés à la représentation statique de la scène. Les objets étant les seuls éléments dynamiques de la scène, ceux-ci doivent être suivis et positionnés dans l'espace pour une reconstruction en direct. En raison de la contrainte temps réel, nous nous tournons vers des représentations pouvant être obtenues rapidement. Pour les utilisateurs, nous créons un *billboard* pour chaque personne de la vidéo afin de représenter leur avatar. Pour l'objet d'intérêt, nous supposons que celui-ci est globalement statique, mais nous souhaitons tout de même qu'il puisse être animé. Pour cela, nous l'avons doté d'un jumeau numérique afin de transmettre ses mouvements relatifs par rapport à sa position à l'initialisation. Cette étape d'actualisation est illustrée figure 5.4.

Implantation

Dans notre implantation, la caméra omnidirectionnelle est disposée sur le lieu commun physique et la scène (environnement, objet d'intérêt et personnes) est construite avec un serveur Python. Pour se télé-immerger sur le lieu, un visiteur se connecte au serveur avec un client Unity pour recevoir les données statiques à l'initialisation et les données dynamiques pour mettre la scène à jour. Celui-ci est immergé en réalité virtuelle dans une représentation 3D de la salle de classe actualisée en temps réel. Les hôtes, les enseignants en particulier, portent un casque de réalité mixte pour voir en surimpression dans la classe les avatars des visiteurs dispersés à travers le lieu. Si un étudiant est à une position dans la reconstruction 3D de la salle, alors un avatar sera affiché à cette même position dans la réalité pour un hôte. Ce mécanisme a été mis en

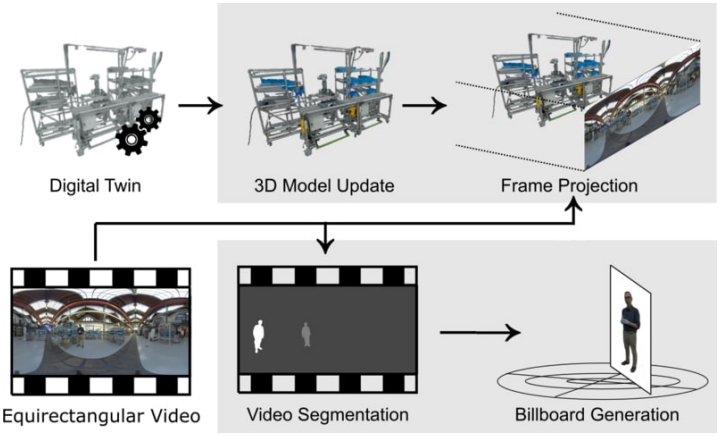


FIGURE 5.4 – Actualisation de la reconstruction 3D 360°. L'objet d'intérêt (boîte du haut) et les personnes (boîte du bas) sont initialisés avec les informations provenant de l'étape d'initialisation.

place avec les avatars Ready Player Me [1] en transmettant en continu la position des étudiants dans la scène au moteur de rendu de l'appareil de réalité mixte (Unity dans notre implantation). Avec les avatars des visiteurs, tous les participants ont conscience de la présence des autres dans la scène et peuvent interagir entre eux.

Dans la suite, nous allons nous concentrer sur la reconstruction 3D des éléments de la scène et détailler comment l'environnement, l'objet d'intérêt et les personnes sont obtenus.

5.2.2 Environnement

L'environnement définit les limites de la scène dans laquelle l'utilisateur se déplace. Les images omnidirectionnelles avec carte de profondeur sont adaptées à la modélisation de l'environnement, mais nous avons vu que cette représentation n'est pas satisfaisante si elle contient des éléments de premier plan (chapitre 4). Dans notre cas où la caméra omnidirectionnelle filme l'objet d'intérêt et des personnes en plus de l'arrière-plan, la carte de profondeur ne peut pas être utilisée directement. L'objectif est donc d'enlever ces éléments de premier plan pour se placer dans le cas où la carte de profondeur est idéale, le cas d'un volume dans lequel un utilisateur peut se déplacer librement sans objets.

La première étape pour créer l'environnement consiste donc à effacer les éléments de premier plan, l'objet d'intérêt et les personnes, de l'image omnidirectionnelle. (Pintore *et al.*, 2022) proposent une solution pour détecter automatiquement les éléments de premier plan sur une image omnidirectionnelle d'intérieur et les effacer grâce à un réseau

1. `https://readyplayer.me`

de neurones. Mais ce réseau de neurones est entraîné sur des images d'intérieurs classiques (environnements domestiques), qui ne se généralise pas à nos lieux. La solution générale consiste plutôt à obtenir un masque de ces éléments et à utiliser un algorithme approprié pour les effacer de l'image. En connaissant le recalage du modèle 3D (décrit sous-section 5.2.3), un masque de l'objet d'intérêt dans l'image équirectangulaire est facilement obtenable. En effet, les coordonnées des fragments qui composent le modèle 3D peuvent être converties en coordonnées sphériques et ainsi retrouver les pixels correspondant sur une projection équirectangulaire (équation 3.1). Afin de créer un masque pour les personnes, une méthode de détection d'objets a été appliquée à l'image omnidirectionnelle. La détection d'objets dans une image omnidirectionnelle est un sujet largement étudié (Yang *et al.*, 2018) et (Fassold, 2019) a observé que le détecteur YOLO entraîné sur des images perspective est robuste aux distorsions équirectangulaires si les éléments ne sont pas proches des pôles ou des bords gauche et droite. Étant donné que nous utilisons une caméra statique sur un sol plat, les personnes présentes dans l'image équirectangulaire tendent à respecter cette contrainte. Nous avons alors utilisé YOLOv8 [2] pour générer les masques des personnes (un masque par personne). En utilisant l'ensemble de ces masques avec la première image de la vidéo, les éléments de premier plan peuvent être effacés à l'aide d'un algorithme d'*inpainting*. Un algorithme d'*inpainting* est un algorithme de traitement d'images visant à effacer des objets non désirés ou à restaurer de manière plausible des parties endommagées ou manquantes (Han et Suh, 2020). Un des usages des algorithmes d'*inpainting* est de supprimer les occultations en utilisant les informations de couleurs aux autres régions pour avoir un arrière-plan sans éléments au premier plan. Cet usage nous permet alors d'effacer les éléments du premier plan de l'image omnidirectionnelle (définis par les pixels du masque) en estimant l'aspect de l'arrière-plan comme s'il n'était pas occulté. Des propositions spécifiquement pour les images omnidirectionnelles ont été soumises (Gkitsas *et al.*, 2021; Han et Suh, 2020) mais aucune ne s'avèrent être capable de fonctionner avec nos données. Nous avons testé plusieurs méthodes d'*inpainting* conçues pour des images perspectives et les résultats les plus convaincants ont été obtenus avec LaMa (Suvorov *et al.*, 2022). Un exemple d'image omnidirectionnelle dont les éléments de premier plan sont effacés avec cette méthode est donné figure 5.5. Avec cette image omnidirectionnelle des éléments d'arrière-plan, la prochaine étape consiste à estimer simplement la carte de profondeur. Cette carte de profondeur est estimée avec notre méthode décrite sous-section 4.1.4. L'avantage de notre méthode comparé aux autres algorithmes testés est qu'elle semble suffisamment robuste pour effectuer une estimation cohérente de la profondeur dans les régions effacées, bien que l'*inpainting* puisse ajouter des artéfacts à

2. `https://github.com/ultralytics/ultralytics`

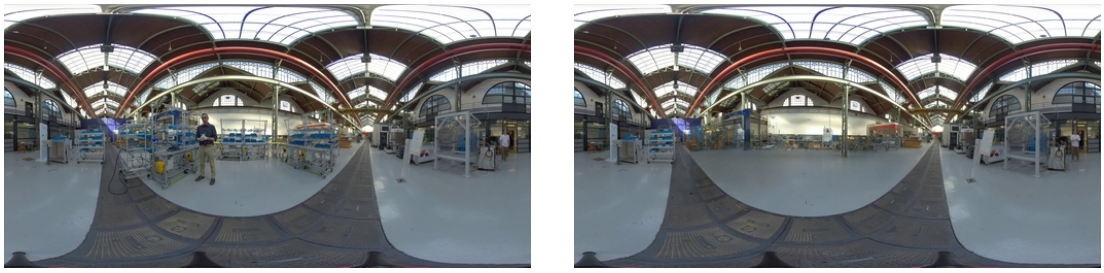


FIGURE 5.5 – Suppression des éléments de premier plan sur une image omnidirectionnelle. Gauche : Image initiale. Droite : *Inpainting*. À noter que dans cet exemple, les éléments de premier plan ont été sélectionnés manuellement.

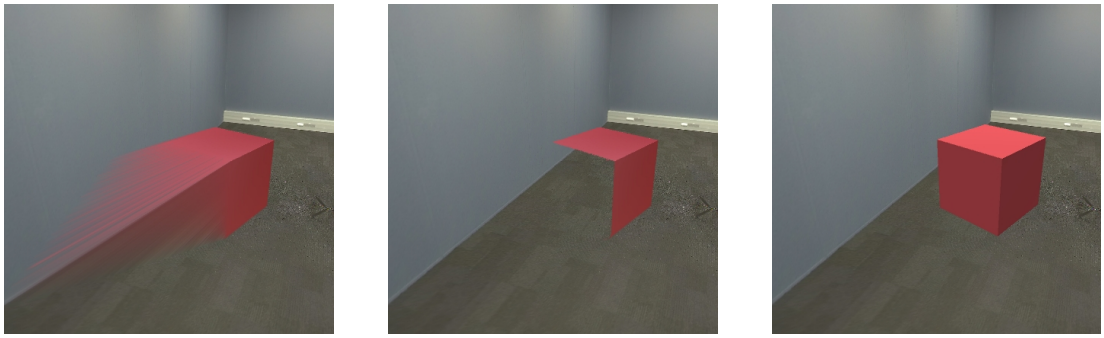


FIGURE 5.6 – Informations géométriques et élément de premier plan. Dans cette scène, les murs et le sol représentent l'arrière-plan et le cube rouge le premier plan. Gauche : La scène est modélisée avec une unique carte de profondeur. L'arrière-plan et le premier plan appartiennent alors à un même maillage 3D connectés par des surfaces fantômes. Milieu : La scène est modélisée par deux cartes de profondeur, une pour l'arrière-plan et une pour le premier plan. On a alors deux maillages 3D distincts, mais le maillage du cube n'est pas complet car seulement la partie du cube visible par la caméra est utilisée. Droite : La scène est modélisée avec une carte de profondeur pour l'arrière-plan et un volume pour le premier plan. Ici, le maillage du cube est complet, indépendamment des régions visibles par la caméra.

l'image. La génération de la carte de profondeur omnidirectionnelle achevant la reconstruction 3D de l'environnement, l'image d'arrière-plan et la carte de profondeur sont alors transmises aux visiteurs.

5.2.3 Objet d'Intérêt

Dans cette version de notre système de télé-immersion, l'objet d'intérêt est l'élément central de la scène de télé-immersion. Idéalement, l'utilisateur interagit avec en se déplaçant autour ou en sélectionnant ses composants. La figure 5.6 compare les différentes informations géométriques pouvant être choisies pour un élément de premier plan. Avec une unique carte de profondeur, il n'y a pas de distinction entre l'arrière-plan et le premier plan, menant à des artéfacts et à l'impossibilité de manipuler indépendamment un

élément de premier plan. En utilisant deux cartes de profondeurs distinctes, l'élément de premier plan est indépendant de l'arrière-plan (il peut être manipulé) mais il n'est alors qu'une surface 3D et non un volume complet, ce qui peut dégrader l'expérience de l'utilisateur s'il désire se déplacer autour. En utilisant un volume, l'élément de premier est un objet naturel de réalité virtuelle et l'utilisateur peut l'observer de tout point de vue. Permettre à l'utilisateur d'inspecter librement l'objet d'intérêt nous impose d'utiliser une représentation volumétrique.

Recalage

Estimer un volume 3D à partir d'un seul point de vue est une tâche complexe. En cohérence avec notre cas d'usage, nous avons supposé qu'un modèle 3D de l'objet d'intérêt devait être connu à l'avance. L'idée est alors de faire comme en réalité augmentée et de superposer un modèle 3D sur son équivalent réel. L'objet d'intérêt est intégré dans la scène simplement en superposant son modèle 3D connu à l'avance sur l'image omnidirectionnelle. (Zanetti *et al.*, 2022) proposent un algorithme pour aligner un modèle 3D sur une image équirectangulaire, basé sur l'analyse de la silhouette. Cependant, leur méthode n'est pas générique : un détecteur spécifique pour le modèle 3D doit être entraîné pour le segmenter sur l'image omnidirectionnelle. Notre but est de pouvoir recaler sur une image omnidirectionnelle un objet quelconque en connaissant seulement son modèle 3D. Une autre possibilité est d'utiliser des marqueurs sur l'objet d'intérêt et de recaler le modèle 3D en fonction de ces marqueurs (López-Cerón et Cañas, 2022). Cependant, il n'existe pas de méthode spécifique pour traiter les marqueurs sur des images omnidirectionnelles, les distorsions de la projection équirectangulaire pourraient perturber les résultats. L'utilisation de marqueurs impose aussi une étape d'instrumentalisation du lieu qu'on souhaite éviter. Enfin, les marqueurs peuvent être difficiles à détecter étant donné la distance qu'il peut y avoir entre l'objet d'intérêt et la caméra omnidirectionnelle. Estimer la pose d'un objet étant équivalent à trouver la pose de la caméra avec la méthode Perspective-n-Point (PnP) (Marchand *et al.*, 2016), nous proposons d'utiliser cette approche basé modèle sans marqueurs pour trouver l'alignement du modèle sur l'image équirectangulaire. Nous nous sommes basés sur SuperPoint (DeTone *et al.*, 2018) pour créer les descripteurs d'une image et SuperGlue (Sarlin *et al.*, 2020) trouver les correspondances entre deux images.

L'approche est résumée par la figure 5.7. La première étape est de déterminer deux caméras perspectives (position, orientation et champ de vision), C_{360} couvrant l'objet d'intérêt sur l'image omnidirectionnelle, et C_{model} qui filme le modèle 3D avec la même pose que sur l'image omnidirectionnelle. Pour trouver ces caméras, l'image omnidirectionnelle est échantillonnée en N images perspectives, et M rendus du modèle 3D sont générés avec des positions de caméras autour du modèle. Ces images sont générées en

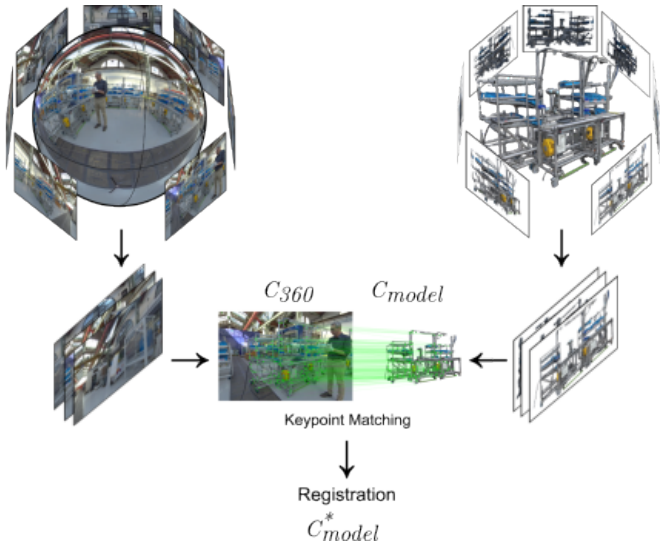


FIGURE 5.7 – Recalage du modèle 3D sur l'image omnidirectionnelle. La meilleure correspondance entre l'échantillonnage de l'image omnidirectionnelle et les rendus du modèle 3D est d'abord recherchée. Le problème PnP est ensuite résolu pour obtenir une correspondance parfaite entre les vues. La différence de pose entre les caméras est alors utilisée pour déterminer la pose du modèle 3D sur l'image omnidirectionnelle.

utilisant le même champ de vision. L'algorithme SuperPoint-SuperGlue est exécuté sur les $N \times M$ paires d'images perspectives, et la paire avec la plus grande correspondance est conservée, C_{360} étant l'image perspective tirée de l'image omnidirectionnelle et C_{model} le rendu du modèle 3D. SuperGlue a l'avantage, par rapport aux autres méthodes testées, d'être plus robuste aux différences de couleurs entre l'image réelle et le modèle 3D. Ainsi, cette correspondance peut être utilisée malgré l'absence de texture réaliste sur le modèle 3D. Bien que les images de la paire ne présentent pas strictement la même vue de l'objet d'intérêt, la correspondance permet d'obtenir une estimation approximative des paramètres C_{model} à utiliser pour avoir la même vue C_{360}. L'utilisation de PnP permet de raffiner la pose de C_{model} pour trouver la caméra perspective donnant la vue la plus proche possible. Les coordonnées 3D des descripteurs produits par Superpoint-Superglue sont d'abord récupérées. Les descripteurs 2D de C_{model} sont projetés sur le modèle 3D par lancer de rayon (l'information de profondeur du modèle permettant de passer de la 2D à la 3D). Ces coordonnées 3D sont alignées sur les coordonnées 2D des descripteurs de C_{360} en résolvant le problème PnP. L'algorithme PnP est initialisé avec la position de C_{model} et exécuté avec RANSAC. La caméra résultante C^*_{model} est la caméra qui couvre le modèle précisément comme C_{360} couvre l'objet d'intérêt. À noter que si les images de C_{model} et C_{360} couvrent déjà l'objet parfaitement de la même manière, alors $C^*_{model} = C_{model}$. La pose précise de la caméra perspective utilisée pour filmer l'objet d'intérêt est alors la translation t et la rotation R entre C^*_{model} et C_{360}.

Enfin, la pose du modèle dans la scène est alors obtenue en appliquant la translation t et la rotation R à partir de l'origine.

Bien que cette méthode permette d'atteindre un alignement assez précis du modèle 3D sur l'objet d'intérêt, elle présente quelques limites. La première est qu'elle ne permet d'aligner qu'une seule instance du modèle 3D sur l'image omnidirectionnelle, car le recalage de plusieurs instances pourrait entraîner des confusions lors de la recherche de correspondances entre des objets similaires. Il est alors impossible d'avoir sur la vidéo deux objets d'intérêt similaires ayant le même modèle 3D. L'autre limite plus problématique est que dans l'idéal, nous aimerions exécuter cet algorithme sur chaque image du flux vidéo afin que les visiteurs aient connaissance en direct de la position de l'objet d'intérêt dans la scène. Malheureusement, cela implique que la recherche de correspondance et la résolution de PnP soit exécutable en temps réel, ce qui n'est pas le cas pour le moment. Nous avons alors dû imposer la contrainte que l'objet d'intérêt devait avoir une position fixe dans la scène pour que le recalage ne soit réalisé qu'une fois à l'initialisation. Si l'objet est dynamique, il est possible d'utiliser une ombre numérique qui, synchronisée avec la vidéo, transmet les mouvements de l'objet d'intérêt.

Actualisation

Les objets pouvaient être dynamiques, l'étape d'actualisation doit mettre à jour le modèle 3D de l'objet d'intérêt. L'algorithme de recalage utilisé étant trop lent pour être exécuté en temps réel, nous utilisons un jumeau numérique pour animer le modèle 3D de l'objet d'intérêt. Un jumeau numérique consiste en un élément physique réel, son jumeau virtuel et une connexion de données entre les deux (Jones *et al.*, 2020b). Ici, nous utilisons specifiquement une ombre numérique où uniquement l'objet physique influence la représentation virtuelle. Dans notre contexte de télé-immersion, les données de l'ombre numérique d'un élément sont une information supplémentaire de la scène. Mais ces informations ne sont pas directement visuelles, elles doivent être interprétées avec un modèle 3D pour reproduire l'apparence de cet élément du côté de l'utilisateur distant. Durant l'étape d'initialisation, un visiteur recevra le modèle 3D de l'objet d'intérêt ainsi que sa position initiale. Les données de l'ombre numérique modifient alors le modèle 3D de l'objet d'intérêt côté client. Par exemple, si l'objet d'intérêt est un cobot, l'ombre numérique contiendra les positions et les vitesses articulaires afin d'animer le modèle 3D tout au long de la vidéo. Les données de l'ombre numérique étant synchronisées avec la vidéo omnidirectionnelle, les mouvements de l'objet d'intérêt sur le modèle 3D et sur la vidéo sont identiques.

Un dernier problème est que le modèle 3D peut ne pas avoir un rendu cohérent avec le reste de la scène. Le modèle 3D pouvant ne pas être réaliste ou bien rendu sans modèle d'éclairage, il peut facilement dénoter avec l'environnement. Pour une meilleure

FIGURE 5.8 – Projection de l'image omnidirectionnelle sur le modèle 3D de l'objet d'intétêt. L'image omnidirectionnelle texture le modèle 3D du cobot aux régions qui ont été visibles par la caméra, la couleur par défaut du modèle est utilisée aux autres régions.

immersion de l'utilisateur, le rendu de l'objet d'intérêt doit être cohérent avec le reste de la scène, mais l'absence d'éclairage ou d'ombres dans le rendu nuit au caractère photoréaliste de la scène alors que c'est un intérêt majeur de la capture d'une scène avec une caméra omnidirectionnelle. Nous souhaitons donc un rendu photoréaliste pour l'objet d'intérêt. Pour obtenir un rendu consistant de l'objet d'intérêt avec l'environnement, l'image courante de la vidéo est utilisée pour texturer le modèle. Les pixels de la vidéo alignés sur le modèle 3D sont projetés sur ce dernier et mettent à jour une texture globale de l'objet d'intérêt. Pour cela, les fragments du modèle 3D de l'objet d'intérêt visibles depuis la position de la caméra sont récupérés grâce à un test d'occultation. Leurs coordonnées sont ensuite converties en coordonnées sphériques. La correspondance simple entre coordonnées sphériques et coordonnées 2D dans l'image équirectangulaire (équation 3.1) permet de déterminer exactement les pixels de texture à appliquer pour chacun des fragments. Ces informations de texture des fragments visibles sont conservées au fil du temps dans une texture globale sous forme d'atlas. Celle-ci est incrémentalement actualisée au fur et à mesure que des nouvelles parties de l'objet d'intérêt deviennent directement visibles par la caméra. La texture globale, initialement remplie avec une couleur par défaut, accumule alors dans le temps des informations de texture sur des régions auparavant occultées lorsque l'objet d'intérêt se déplace. Si une région n'est jamais vue par la caméra, les pixels correspondants dans la texture globale conservent la couleur par défaut (celle-ci peut être la couleur par défaut du modèle 3D). Ce rendu de texture persistant dans le temps permet de rendre le modèle photoréaliste sur le plus de régions possible. TODO VOIR FIGURE figure 5.8.

FIGURE 5.9 – Avatars monocaméra. Ces représentations sont reconstruites à partir d'une seule vue de face d'une personne. Gauche : PIFuHD (Saito *et al.*, 2020). Droite : *Billboard*.

5.2.4 Personne

En ajoutant un avatar pour chaque personne sur le lieu commun, notre représentation 3D 360° est complète avec l'environnement et l'objet d'intérêt. De nombreuses représentations sont possibles pour incarner un utilisateur sous forme d'avatar : avec un avatar 3D de l'utilisateur connu à l'avance (scan a priori) ou en créant son modèle à la volée. En connaissant les avatars 3D à l'avance, il est possible d'estimer la posture de chaque personne grâce des méthodes d'estimation de pose humaine pour images omnidirectionnelles (Aso *et al.*, 2021; Shere *et al.*, 2019) et d'ajouter des avatars qui reproduisent les mêmes poses dans la scène (Dupont de Dinechin et Paljic, 2019). Cependant, contrairement à l'objet d'intérêt, nous faisons l'hypothèse que les utilisateurs n'ont pas de scan 3D connu à l'avance prêt à être réutilisé. Nous supposons aussi qu'il est plus agréable pour les visiteurs de voir les hôtes avec une représentation qui leur est propre (avatar personnalisé) plutôt qu'un avatar générique. L'idée est alors, pour chacun des hôtes, de le détecter précisément sur le flux omnidirectionnel, et de lui générer une incarnation à partir de cette vue. La difficulté principale pour obtenir une représentation 3D acceptable d'une personne vient de l'utilisation d'une unique caméra omnidirectionnelle statique. En effet, la reconstruction d'un avatar volumétrique 3D à partir d'un unique point de vue sans informations géométriques est un problème ambigu. À notre connaissance, seul Monoport (Li *et al.*, 2020) offre une création en temps réel d'un avatar à partir d'un seul flux vidéo en direct. L'algorithme a la capacité de reconstituer des régions de l'avatar qui ne sont pas directement vues par la caméra en se reposant sur PIFu (Saito *et al.*, 2019), un réseau de neurones pour la génération d'avatar 3D à partir d'une image (figure 5.9 gauche). Cependant, ce réseau de neurones a l'inconvénient d'être sujet à des erreurs de reconstruction susceptibles de dégrader le

sentiment de présence de l'utilisateur.

Notre proposition est de représenter les personnes par des *billboards*. Un *billboard* (Fourquet *et al.*, 2007; Horry *et al.*, 1997), aussi appelé imposteur (Livatino, 2007; Hamill *et al.*, 2005), est un plan 2D positionné dans l'espace 3D sur lequel est projetée une image ou une vidéo. Généralement utilisée pour représenter les objets ambiants, cette représentation est pertinente lorsqu'une reconstruction 3D d'un objet ne peut pas être réalisée par manque de point de vue autour de celui-ci. Dans notre contexte, avec juste une vue de face des hôtes, nous projetons les vidéos des personnes sur des *billboards* pour représenter leurs avatars (figure 5.9 droite). En termes de représentation stratifiée (section 5.1), cela revient à utiliser une couche par personne. Cette forme d'avatar n'est donc ni volumétrique ni en relief. Bien qu'il soit possible d'ajouter du relief avec une carte de profondeur, nous avons choisi d'utiliser un plan plat pour des raisons de performance, mais aussi car nous supposons que l'ajout de relief au *billboard* est marginal sur la perception de l'avatar. Notre choix pour le *billboard* plutôt qu'un avatar volumétrique est appuyé par des expériences montrant que cette représentation est plus confortable pour les utilisateurs qu'une reconstruction bruitée (Debarba *et al.*, 2022) et équivalente à un avatar 3D si l'utilisateur fait face au *billboard* (Cho *et al.*, 2020). L'objectif de notre algorithme est alors de générer des *billboards* pour chacune des personnes uniquement à partir du flux vidéo omnidirectionnel. L'avantage de notre approche est qu'elle crée en temps réel une représentation d'un objet dans la vidéo de manière générique à partir du moment où l'objet est détecté lors de l'initialisation. En effet, notre approche ne repose sur aucun mécanisme spécifique au traitement des humains comme le suivi de point clé humain. Il est donc possible de créer sans distinctions un *billboard* pour n'importe quel objet de la scène. Cependant, la création d'un *billboard* pour un objet repose sur sa détection dans l'image. La segmentation entre arrière-plan et premier plan restant imprécise aujourd'hui, nous utilisons les *billboards* uniquement pour les personnes sur le flux vidéo. Un autre avantage de cette représentation est qu'elle consiste à transmettre une vidéo augmentée de quelques valeurs scalaires, les techniques de communication vidéo sur le réseau peuvent alors être exploitées efficacement. Les avatars sont alors transmis avec une bande passante réduite (comparé à un avatar volumétrique) et le système de télé-immersion est alors déployable sur des ordinateurs avec une configuration standard. Finalement, notre approche pour créer un *billboard* d'une personne se décompose en deux étapes : l'extraction d'une vue de face de la personne sous forme d'image perspective et la création du plan dans l'espace 3D.

Extraction Vue Perspective

La première étape pour créer les *billboards* consiste à extraire les vues de face de chaque personne à partir d'une image omnidirectionnelle. De cette étape résulte une

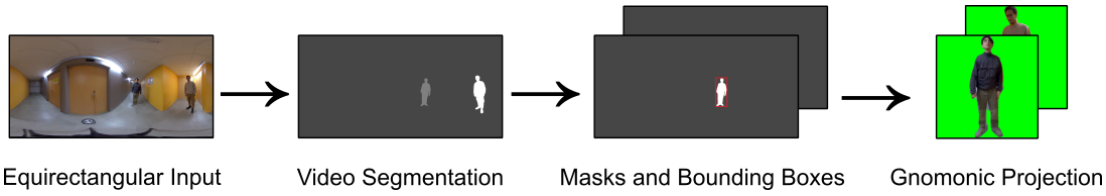


FIGURE 5.10 – Extraction de l'image omnidirectionnelle des vues centrées sur les personnes. L'image omnidirectionnelle est segmentée pour avoir un masque omnidirectionnel pour chacune des personnes sur l'image. Une boite englobante de la personne est calculée sur chaque masque. La valeur d'une boite englobante permet de calculer une projection gnomonique centrée sur la personne. Dans l'image finale, les pixels n'appartenant pas au masque de la personne sont remplacés par une couleur par défaut (ici en vert).

vue perspective pour chaque personne (centrée sur elle). La méthode pour obtenir ces vues est illustrée figure 5.10. Cette extraction débute par une segmentation vidéo de l'ensemble de la vidéo équirectangulaire pour isoler les personnes des autres éléments. Pour cela, nous utilisons XMem (Cheng et Schwing, 2022) qui segmente la vidéo en restant robuste aux occultations. Cette segmentation vidéo résulte en un masque omnidirectionnel des personnes présentes dans les images de la vidéo omnidirectionnelle. Contrairement à des détections d'humains sur chaque image de la vidéo de manière indépendante, la segmentation conserve une cohérence temporelle entre les masques prédits. Les méthodes de segmentation vidéo doivent généralement être initialisées avec un masque pour déterminer les éléments à suivre. Notre segmentation vidéo est alors initialisée avec le masque de la détection humaine de la reconstruction statique (sous-section 5.2.2). Avec le masque obtenu par la segmentation, il est possible d'identifier à quelle personne appartient un pixel ou si c'est un pixel d'arrière-plan. Le principal intérêt de traiter la totalité de l'image équirectangulaire est que la segmentation vidéo permet de créer tous les *billboards* en une seule passe (au lieu de générer les vues perspectives qu'il faudrait ensuite individuellement segmenter). Sachant que la segmentation vidéo peut prendre du temps, la mise à jour du masque dans notre pipeline est effectuée sur une image à résolution réduite afin d'atteindre un traitement en temps réel. Après la prédiction sur l'image réduite, le masque est rétabli à la résolution d'origine par interpolation bicubique. Toujours pour accélérer le temps de calcul, tous les traitements d'images possibles sont réalisés sur GPU avec CuPy [3].

À partir de cette segmentation, une image perspective est extraite de l'image équirectangulaire pour chacune des personnes sur la vidéo. Pour passer du masque équirectangulaire à une image perspective centrée sur une personne, une projection gnomonique est calculée pour chaque personne suivie. Les paramètres de la projection gnomonique sont estimées à partir d'une boîte englobante alignée sur les axes ($AABB$) des pixels du

3. `https://cupy.dev`

masque, de centre (i_{bb}, j_{bb}) et de taille (w_{bb}, h_{bb}). Le centre de la boite englobante permet de calculer le centre de la projection gnomonique (φ, θ) à l'aide de l'équation 3.1. La taille de la boite englobante permet de calculer le champ de vision f de la projection gnomonique avec la formule suivante :

$$\begin{aligned} f_h &= 2\pi \frac{w_{bb}}{w} \\ f_v &= \pi \frac{h_{bb}}{h} \\ f &= \max(f_h, f_v) \end{aligned} \tag{5.1}$$

où w et h sont la largeur et la hauteur de l'image équirectangulaire. Cette formule utilise aussi la correspondance entre une taille en pixels et un angle dans la projection équirectangulaire. f_h correspond au champ de vision horizontal minimal pour que la personne soit capturée en entier en largeur, et f_v au champ de vision vertical minimal pour que la personne soit capturée en entier en hauteur. Le champ de vision d'une projection gnomonique étant le même verticalement et horizontalement, le maximum entre f_h et f_v est choisi pour que la personne soit capturée en entier sur l'image perspective. Avant d'effectuer la projection gnomonique pour une personne, l'image équirectangulaire d'entrée est masquée à l'aide de la segmentation prédite, et les pixels situés en dehors du masque sont remplacés par une couleur par défaut. Ainsi, l'image perspective résultante est une image carrée, où les pixels de la personne (prédite par le masque) garde leur couleur et les autres pixels ont une couleur par défaut.

Création du Plan

Après avoir obtenu une image à afficher sur le plan, l'étape finale consiste à créer le plan dans l'espace. La création du plan nécessite de trouver deux paramètres : sa position et sa taille. Positionner le plan nécessite d'estimer la distance entre la caméra et la personne sur le plan. En l'absence d'information de profondeur, cette distance est approximée. Les informations de pose d'une personne peuvent aider à approcher la distance à la caméra (Mehta *et al.*, 2017) mais cette formule ajouterait une étape d'estimation de pose d'humain superflue. La distance entre la caméra et un objet peut aussi être calculée en fonction de la taille d'un objet dans l'image (Sun et Zollmann, 2022) mais cette approche manque de fiabilité. Nous privilégions une méthode d'évaluation des distances se fondant sur les propriétés de l'image équirectangulaire avec la connaissance a priori de la hauteur de la caméra. La distance ρ entre la caméra et une

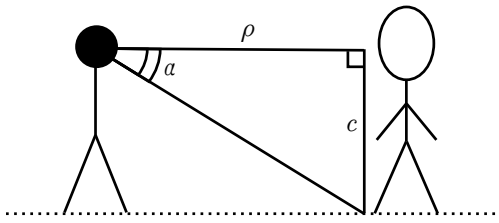


FIGURE 5.11 – Estimation de la distance entre la caméra et un objet.

personne est alors calculée à l'aide de la formule proposée par (Mazzola *et al.*, 2021) :

$$\begin{aligned} \alpha &= \frac{\pi}{2} \frac{y_{bb} + \frac{h_{bb}}{2} - \frac{h}{2}}{\frac{h}{2}} \\ \rho &= \frac{c}{\tan \alpha} \end{aligned} \tag{5.2}$$

où c est la hauteur de la caméra. La justification de la formule est donnée figure 5.11. La formule suppose que l'objet touche le sol, que le sol est plat et que la hauteur de la caméra est connue, hypothèses valables dans notre contexte. Le point de contact entre le sol et l'objet et l'horizon de la caméra permettent de définir l'angle α. Pour calculer la distance ρ, on suppose que le point de contact entre une personne et le sol correspond au pixel du masque équirectangulaire avec la valeur en ordonnée la plus grande, c'est-à-dire $y_{bb} + \frac{h_{bb}}{2}$ dans un repère où $y = 0$ est le haut de l'image. L'angle α est alors la conversion en angle de la distance en pixels entre cette hauteur et le milieu de l'image équirectangulaire. Connaissant ces valeurs, la distance ρ est calculée simplement par trigonométrie.

À partir de la distance, la taille du plan et sa position peuvent être déterminées. La situation pour la création du plan est présentée figure 5.12. La taille d'un plan est estimée avec la formule suivante :

$$s = 2\,\rho \tan \frac{f}{2} \tag{5.3}$$

Étant donné que les champs de vision horizontal et vertical sont les mêmes avec une projection gnomonique, le plan doit être un carré pour éviter les distorsions. Ainsi s est la hauteur et la largeur du plan. On considère que la projection gnomonique permet d'obtenir une vue de face de la personne, comme si une caméra perspective avait été placé en face d'elle. En positionnant les plans toujours face à la caméra sans inclinaison (les plans perpendiculaires au sol), cette formule se déduit par trigonométrie, et indépendamment du fait que le champ de vision f soit f_h ou f_v. D'autres estimations de s sont aussi possibles, mais au prix d'un traitement différent selon que le champ de vision soit f_h ou f_v.

Enfin, les coordonnées cartésiennes du centre du *billboard* (x, y, z) sont estimées à

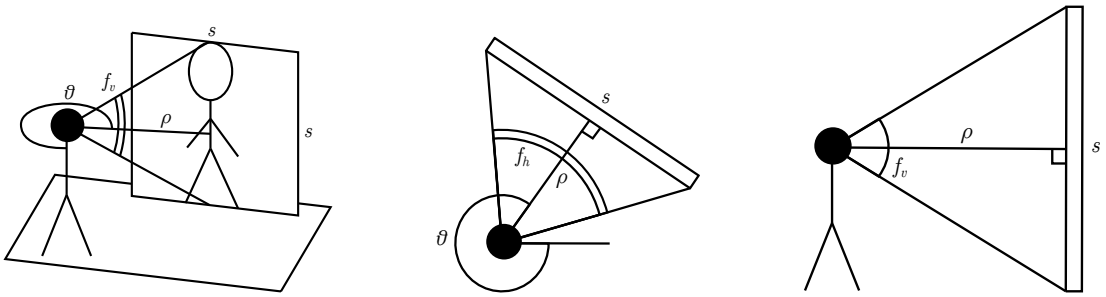


FIGURE 5.12 – Création du plan pour un *billboard*. Gauche : Vue 3D. Milieu : Vue du haut. Droite : Vue de profil.

l'aide des formules suivantes :

$$\begin{aligned} x &= \rho \cos\theta \\ y &= 2\,\rho \tan\frac{f_v}{2} \\ z &= \rho \sin\theta \end{aligned} \tag{5.4}$$

Les coordonnées x et z sont simplement déduits à partir de l'angle horizontal θ correspondant à l'orientation de la personne autour de la caméra. Pour calculer y, on doit connaître la hauteur de la personne sur l'image perspective. Cette hauteur s'obtient comme pour la taille du plan s, mais en utilisant spécifiquement le champ de vision vertical f_v. En effet, si $f = f_h$, la hauteur de la personne sur l'image n'est pas la hauteur de l'image, et donc s n'est pas une bonne estimation de y. L'estimation de y avec f_v permet donc d'avoir une bonne hauteur de plan indépendamment de la valeur de f. À noter que dans le cas où le champ de vision est f_h, le plan est enfoncé dans le sol pour que la personne sur l'image semble bien touché le sol. L'avantage de la segmentation vidéo est que la cohérence temporelle due au modèle de suivi donne des valeurs de θ, f_h et f_v lissées sur le temps. Cette cohérence donne un mouvement lisse des personnes dans l'espace, minimisant les effets de déplacements saccadés. Mais cet effet de déplacements saccadés est tout de même présent lorsqu'une personne se rapproche ou s'éloigne de la caméra à cause de la méthode d'estimation de distance. En effet, ρ étant calculé à partir du point de contact entre une personne et le sol et que ce point n'évolue pas de manière progressive quand la personne se rapproche ou s'éloigne mais par saut, la distance n'est pas lisse dans le temps. Même s'il peut être observable, nous pensons qu'un modèle statistique approprié peut corriger ce problème pour lisser les valeurs de ρ dans le temps.

Une fois cette étape terminée, les vidéos et les paramètres calculés sont transmis aux visiteurs pour afficher les hôtes dans la scène. Les *billboards* sont rendus sous forme d'objet 3D par incrustation pour rendre les pixels d'arrière-plan transparent. Un exemple de

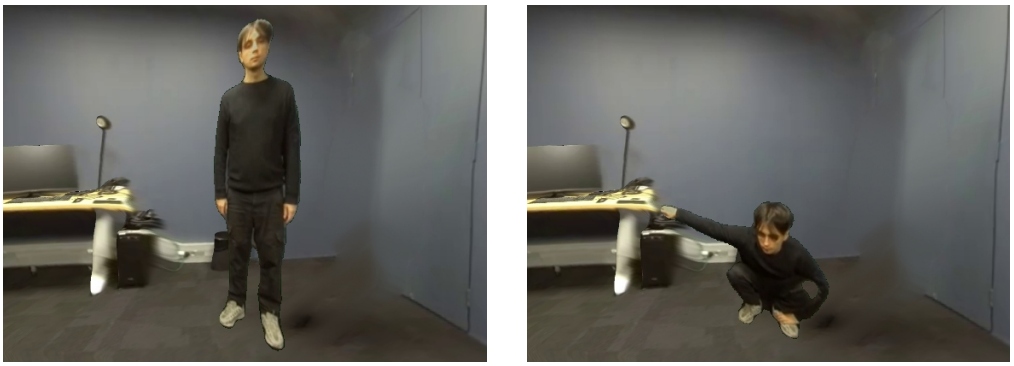


FIGURE 5.13 – Résultats de la méthode de génération de *billboards*. Les *billboards* sont intégrés dans un environnement 3D obtenu à partir de la vidéo omnidirectionnelle.

billboard avec l'environnement obtenu à partir d'une vidéo omnidirectionnelle est donné figure 5.13. À noter que l'orientation des *billboards* est fixe, toujours face à la caméra et ne s'orientent pas vers l'utilisateur. Cette orientation fixe a été décidée pour maintenir une cohérence entre un hôte et le reste de la scène, par exemple s'il manipule l'objet d'intérêt. Les visiteurs voient donc tous les mêmes *billboards* des hôtes.

5.2.5 Évaluation

Avant d'expérimenter avec des utilisateurs, nous évaluons notre proposition de téléimmersion sur des critères techniques. L'objectif de l'évaluation est de justifier que le système peut être utilisé en pratique, c'est-à-dire qu'il peut atteindre des temps d'interaction et une qualité visuelle raisonnable. Pour s'assurer que notre système est interactif, nous mesurons les fréquences d'images en fonction de la résolution de la vidéo et du nombre de personnes suivies. En ce qui concerne le rendu, un facteur important est la qualité du détourage des personnes suivies. La qualité de la segmentation vidéo a également été évaluée en fonction de la résolution de la vidéo et de la distance par rapport à la caméra. Pour nos évaluations, les calculs ont été effectués sur un serveur équipé d'un GPU NVIDIA GeForce RTX 3090.

Évaluation des Performances

Pour vérifier si les éléments dynamiques peuvent être rendus en temps réel, la durée de l'ensemble du processus de génération du *billboard* a été mesurée, depuis la segmentation de la vidéo au calcul de la position 3D. Trois résolutions de vidéo ont été testées : basse (400×200), moyenne (800×400) et haute (1600×800) afin de déterminer la résolution maximale atteignable en temps réel. Le changement de résolution n'a pas d'effet sur la qualité de l'image, mais affecte la qualité du détourage des personnes. De plus, une à trois personnes ont été suivies afin de déterminer si le système est suffisamment

TABLE 5.1 – Fréquence d'images moyenne en fonction de la résolution de l'image équirectangulaire et du nombre de personnes suivies

Résolution		Fréquence d'images
Basse (400 × 200)	1 personne	38.4 fps
	2 personnes	23.9 fps
	3 personnes	17.6 fps
Moyenne (800 × 400)	1 personne	30.5 fps
	2 personnes	18.9 fps
	3 personnes	13.8 fps
Haute (1600 × 800)	1 personne	11.2 fps
	2 personnes	7.8 fps
	3 personnes	5.9 fps

rapide pour traiter plusieurs personnes. Les fréquences d'images obtenues sont indiquées dans le tableau 5.1. Ces résultats révèlent que l'interactivité du système n'est actuellement possible qu'en résolution basse et moyenne avec une seule personne. Le coût supplémentaire lors du suivi de plusieurs personnes est principalement dû au fait que le positionnement des *billboards* dans l'espace 3D est effectué de manière séquentielle pour chaque personne. Une parallélisation pourrait conduire à une augmentation de la fréquence d'images lorsqu'il y a plusieurs personnes. Aussi, aucun effet de la résolution du *billboard* en sortie n'a été observé sur la fréquence d'images.

Pour s'assurer que la représentation proposée est diffusée sans charge excessive sur le réseau, nous avons également évalué la bande passante utilisée par notre approche. L'environnement est transmis en HTTP, les vidéos des *billboards* en RTSP et les poses des *billboards* en UDP. La résolution de la vidéo du *billboard* pour une seule personne est de 800×800 pixels. Si plusieurs personnes sont présentes dans la scène, leurs images sont assemblées en une image commune plus grande afin de n'utiliser qu'un seul flux RTSP. À titre de comparaison, la vidéo omnidirectionnelle originale utilisée pour générer la scène est également diffusée en flux RTSP avec une projection équirectangulaire d'une résolution de 2048×1024 pixels. Nous constatons que l'utilisation de la bande passante de notre approche est proche de celle utilisée pour transmettre la vidéo omnidirectionnelle, à l'exception du démarrage où notre approche reçoit une quantité de données plus importante pour obtenir la reconstruction 3D statique. La latence observée est d'environ 1.5 seconde, ce qui semble acceptable pour une interaction entre des visiteurs et des hôtes. Ces résultats valident notre objectif de navigation libre en direct avec une utilisation du réseau équivalente à celle de la visioconférence.

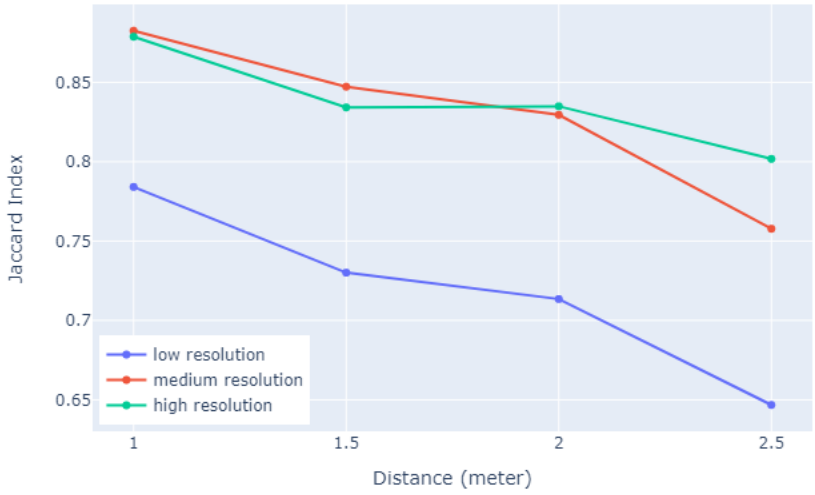


FIGURE 5.14 – Qualité du détourage d'une personne en fonction de la distance à la caméra et de la résolution d'entrée.

Évaluation de la Qualité Visuelle

Un autre facteur important pour que les utilisateurs adhèrent à notre approche est la qualité visuelle de notre représentation. Un point intéressant est alors d'évaluer la qualité des *billboards* et vérifier s'il n'y a pas d'erreurs flagrantes dans la segmentation des personnes sur la vidéo. Ces erreurs peuvent conduire à altérer l'incarnation en *billboard*, ce qui dégraderait la communication non-verbale. L'expérience vise à déterminer l'évolution de la qualité du détourage en fonction de la distance à la caméra afin de fournir des informations sur les bonnes pratiques avec le système. Nous formulons l'hypothèse selon laquelle plus la résolution est élevée, plus la qualité du contour est élevée, et que la qualité du contour diminue lorsque la distance à la caméra augmente.

Pour évaluer la qualité du détourage, nous tournons d'abord plusieurs vidéos avec une personne à différentes distances de la caméra (1 m, 1.5 m, 2 m et 2.5 m). Cette personne prend un certain nombre de poses différentes qui sont les mêmes dans toutes les vidéos. Dans ces vidéos, certaines images clés ont été sélectionnées pour créer manuellement des masques de référence. Ces masques de référence, utilisés comme vérité de terrain, ont été obtenus à l'aide d'un outil de segmentation interactif (Sofiiuk *et al.*, 2022) en sélectionnant manuellement la personne sur l'image équirectangulaire. Ensuite, la segmentation est effectuée sur les vidéos, en prenant comme entrée les résolutions basse, moyenne et haute. Les résultats de la segmentation vidéo avec les différentes résolutions

de ces images clés ont été enregistrés et comparés aux masques de référence à l'aide de l'indice de Jaccard (*Intersection over Union*). La moyenne des indices de Jaccard des images clés pour chaque distance est calculée pour obtenir le graphe figure 5.14. Sur la base de ces résultats, nous observons que la qualité du contour se dégrade à mesure que la distance par rapport à la caméra augmente, confirmant ainsi notre hypothèse. À noter que les valeurs des indices de Jaccard sont élevées car l'image a une résolution de 2048×1024 alors que la segmentation de la personne n'occupe qu'une petite partie de cette image. Cependant, la métrique semble indiquer que les résolutions moyenne et haute ont à peu près la même qualité de contour alors que celle de la basse résolution est bien inférieure. Cette observation suggère qu'une résolution moyenne peut être utilisée pour des raisons de performance sans perte de qualité. Nous remarquons également sur les masques obtenus que toutes les résolutions ont tendance à perdre la trace des bras et des mains après 1.5 m, ce qui est préjudiciable pour la communication non verbale.

Ces résultats nous permettent de conclure que pour une utilisation interactive de notre système avec une représentation en *billboard* agréable, la meilleure configuration est d'utiliser une résolution moyenne (capable d'atteindre des performances temps réel) avec une seule personne positionnée entre 1 m et 1.5 m de la caméra.

5.2.6 Limites

Dans cette section, nous avons présenté la version définitive de notre système de télé-immersion. Cette version répond à nos besoins : l'utilisateur peut se déplacer librement dans un lieu dynamique et les objets sont représentés comme des éléments indépendants. Néanmoins, de nombreuses limites restent encore à surmonter. Une première est que les technologies que nous utilisons (segmentation, détection, *inpainting*) sont conçues pour fonctionner sur des images perspectives. Notre système manque alors de robustesse lorsque les propriétés sphériques des images équirectangulaires doivent être exploitées. Par exemple, un hôte passant du bord gauche de l'image au bord droit (ou vice versa) peut ne pas être correctement suivi. La représentation des objets (objets d'intérêt et avatar) reste encore à approfondir. Le problème majeur s'avère être l'incarnation sous forme de *billboard*. Il est possible que l'absence de 3D amène une diminution du réalisme si le visiteur ne fait pas face à un hôte et qu'un avatar 3D volumétrique puisse alors être préféré (Cho *et al.*, 2020). Pour l'objet d'intérêt, la manière de le texturer n'est pas totalement satisfaisante au regard de la possibilité de se déplacer librement autour car seules les régions ayant été visibles ont une texture photoréaliste sur le modèle 3D. Ce manque d'informations dégrade l'expérience de l'utilisateur et l'incite à ne rester qu'en face de l'objet d'intérêt. Aussi, l'animation de l'objet d'intérêt requiert une ombre numérique alors qu'une meilleure solution serait de pouvoir le suivre en temps réel le long de la vidéo. Enfin, un problème inattendu plus général de notre système est le manque de

fidélité dans l'échelle de l'environnement. Dans notre système, avoir une reconstruction à l'échelle est critique car si un visiteur se déplace à une certaine distance de la caméra dans la scène, l'hôte doit voir ce visiteur à la même distance. Si cette distance n'est pas la même, l'hôte verra le visiteur à une position qui n'est pas la sienne dans la scène. Par exemple, si l'échelle est trop petite, le visiteur sera bien plus loin de la position de la caméra dans la reconstruction 3D que l'hôte ne le verra. Pour que ces positions dans la scène et dans le lieu physique soient les mêmes, la géométrie et l'échelle doivent être estimées finement. Or, nous n'avons pas trouvé de méthode pour estimer l'échelle automatiquement pour que ces positions semblent bien se superposer.

Des solutions sont tout de même identifiées pour augmenter la qualité visuelle de la scène. Une première solution est d'utiliser les techniques de reconstruction 4D pour construire les objets le long la vidéo. L'idée est de segmenter les différents objets sur la vidéo omnidirectionnelle et d'accumuler des informations géométriques pour avoir un modèle 3D de plus en plus correct. Cette approche est plus facilement mise en pratique pour la reconstruction d'avatar étant donné que la détection de personne sur une image est largement étudiée. Pour l'objet d'intérêt, il est possible de continuer à utiliser un modèle connu à l'avance (comme il est plus difficile de le détecter) et la tâche serait de le suivre directement sur la vidéo en temps réel. Pour texturer totalement l'objet d'intérêt, une possibilité serait de trouver comment propager la texture des zones visibles aux zones occultées. La solution pour halluciner des textures existe déjà pour les avatars volumétriques afin de déterminer la couleur de régions qui ne sont pas directement visibles (Xu et Loy, 2021; Saito *et al.*, 2019). Sans modèle à l'avance, la reconstruction 3D à partir d'une unique image peut aussi être envisagée, même si aujourd'hui les méthodes fonctionnent uniquement pour des catégories précises d'objets (Monnier *et al.*, 2022; Zhang *et al.*, 2022a; Xu *et al.*, 2021b). Même s'il est désirable de pouvoir s'en passer, l'utilisation d'un jumeau numérique de l'objet d'intérêt reste intéressant. En passant d'une ombre numérique à un jumeau numérique complet, les étudiants à distance interagissent directement avec l'objet d'intérêt physique à travers le modèle 3D, ce qui permettrait d'étendre le système de télé-immersion à d'autres usages comme la réalisation de travaux pratiques en ligne. Enfin, la représentation de l'environnement peut aussi être sujette à améliorations. En nous basant encore sur une carte de profondeur pour l'environnement, notre approche de télé-immersion n'est toujours pas appropriée pour les lieux extérieurs. Les représentations stratifiées apparaissent alors être plus intéressante et leurs études pourraient permettre d'avancer de nouvelles représentations avec une parallaxe plus large.

5.3 Conclusion

Ce chapitre conclut le développement de notre système de télé-immersion. Cette version définitive est spécifiquement conçue pour notre cas d'usage d'enseignement à distance. Les étudiants à distance en réalité virtuelle naviguent librement dans la salle de classe pour étudier un objet d'intérêt. Le professeur sur site, qui capture le lieu avec une caméra omnidirectionnelle statique, voit les étudiants dans la salle de classe sous forme d'avatars grâce à un casque de réalité mixte. Cette version est la plus proche du système télé-immersion idéal présenté section 1.3. Notre approche se base sur une représentation 3D 360° du lieu d'intérêt qui répond aux besoins de notre cas d'usage. La représentation combine, une image omnidirectionnelle avec carte de profondeur pour avoir un environnement dans lequel le visiteur navigue librement, un modèle 3D pour avoir un objet d'intérêt avec lequel le visiteur peut interagir (se déplacer autour, le sélectionner ou le manipuler) et des avatars en *billboards* pour les hôtes afin que le visiteur puisse communiquer avec eux. La génération de cette représentation en direct étant cruciale pour notre cas d'usage, nous proposons une stratégie pour la reconstruire en temps réel en traitant séparément les éléments statiques et dynamiques. L'environnement, statique, est reconstruit à l'initialisation grâce à l'*inpaiting* et l'estimation de profondeur omnidirectionnelle. La pose du modèle 3D de l'objet d'intérêt est estimée avec une méthode de recalage sur la vidéo omnidirectionnelle, mais cette méthode étant trop coûteuse pour un traitement temps réel, ce recalage est effectué uniquement à l'initialisation. Pour animer l'objet d'intérêt, une ombre numérique est alors utilisée, synchronisée avec le flux vidéo. Enfin, les *billboards* sont créés facilement grâce à une méthode légère basée sur une segmentation vidéo du flux.

La représentation 3D 360° proposée permet d'accomplir notre besoin de navigation libre. Nous avions expliqué au chapitre précédent que l'image omnidirectionnelle avec carte de profondeur convenait généralement pour des lieux simples sans objets au premier plan (typiquement une salle vide quand tout est visible). Les méthodes de traitement d'images dans l'état de l'art et notre propre proposition pour éliminer le premier plan sur notre image omnidirectionnelle semble aller dans le sens que l'acquisition de l'environnement sans les objets sera de plus en plus aisée. Toutefois, ces éléments de premier plan doivent être modélisés en 3D et incorporés dans l'environnement pour une représentation complète de la scène. Une question demeure alors : quelle représentation utiliser pour les objets ? À partir d'une image omnidirectionnelle, seulement une vue de face de l'objet est disponible. La génération d'un modèle 3D à partir de cette unique image reste aujourd'hui un problème technique difficile, et la question est alors de savoir quelle représentation crédible pour un visiteur peut être obtenue. Dans notre représentation 3D 360°, nous avons supposé pour les personnes que le *billboard* était une représentation acceptable. Dans le chapitre suivant, nous vérifions cette hypothèse :

nous discutons qu'avec notre représentation 3D 360°, les visiteurs se sentent effectivement sur le lieu commun. En particulier que le *billboard* est bien un choix raisonnable pour modéliser les hôtes.

Chapitre 6

Présence en Télé-Immersion 3D 360°

La version définitive de notre système de télé-immersion 3D 360° achevée, il est essentiel d'évaluer son apport. Ce système se base sur une représentation 3D 360° qui combine carte de profondeur omnidirectionnelle pour l'environnement, modèle 3D aligné pour l'objet d'intérêt et *billboards* pour les personnes. L'immersion et l'interaction sont plus grandes qu'avec la télé-immersion 360° actuelle qui immerge simplement les utilisateurs dans un flux 360° filmé en direct par la caméra. Pour argumenter l'apport de notre système par rapport à la télé-immersion 360° classique, le ressenti des utilisateurs doit être recueilli. Un aspect intéressant à explorer concernant un système de télé-immersion est le sentiment de présence qu'il induit aux utilisateurs, nous cherchons donc particulièrement à mesurer ce sentiment chez les utilisateurs de notre système. Nous faisons l'hypothèse que notre système de télé-immersion est supérieur car il induit une plus grande présence que la télé-immersion 360° classique. Dans ce chapitre, nous rapportons une expérience visant à vérifier cette hypothèse.

6.1 Présence en Télé-Immersion

Au cours des chapitres précédents, nous avons exploré des approches pour progresser vers un système de télé-immersion dans lequel plusieurs visiteurs peuvent se déplacer librement sur un lieu capturé par une caméra omnidirectionnelle. Dans sa forme définitive, notre système repose sur une représentation 3D 360° constituée d'une image omnidirectionnelle avec carte de profondeur, d'un modèle 3D et de *billboards* (sous-section 5.1.2). Afin d'attester de l'intérêt de notre système de télé-immersion et de la représentation 3D 360°, nous avons conduit des évaluations objectives. Nous avons évalué en particulier les performances du système ainsi que la qualité du détourage des *billboards* (sous-section 5.2.5). Un système de télé-immersion peut aussi être évalué sur

d'autres critères objectifs comme la compression des données, la latence ou la qualité d'images (Wang *et al.*, 2023b; Siemonsma et Bell, 2022; Lawrence *et al.*, 2021). Mais recueillir les avis subjectifs d'utilisateurs est également crucial pour garantir que le système est utilisable en pratique, par exemple en identifiant les préférences entre différents systèmes ou en comparant avec une situation en face à face (Alvarez *et al.*, 2022; Zhang *et al.*, 2022b; Lawrence *et al.*, 2021; Pejsa *et al.*, 2016). L'utilisabilité d'un système, mesurée avec le *System Usability Scale* SUS (Brooke, 1996), ou la charge de travail mesurée avec le NASA-TLX (Hart et Staveland, 1988) sont des métriques fréquentes lors de l'évaluation d'un système de télé-immersion (Luo *et al.*, 2023; Kolkmeier *et al.*, 2018). Les systèmes élaborés pour des usages précis sont aussi évalués sur l'accomplissement d'une tâche, par exemple en tenant compte du temps passé (Teo *et al.*, 2019; Lee *et al.*, 2018; Poelman *et al.*, 2012). Pour notre cas d'usage d'enseignement à distance, des critères tels que l'engagement des étudiants ou la rétention d'information pourraient être évalués. Ces valeurs sont d'ailleurs plus élevées avec une vidéo omnidirectionnelle qu'avec une vidéo classique (Chan *et al.*, 2021; Harrington *et al.*, 2018). Mais, une donnée pertinente à retrouver sur un système de télé-immersion, générique au cas d'usage, est sa capacité à susciter un sentiment de présence chez l'utilisateur (Alvarez *et al.*, 2022; Rhee *et al.*, 2020; Teo *et al.*, 2019; Kolkmeier *et al.*, 2018; Pejsa *et al.*, 2016).

L'étude de la présence, le sentiment d'un utilisateur d'être réellement dans un autre lieu que son lieu physique, est courante en télé-immersion. Une hypothèse répandue est qu'une augmentation du sentiment de présence sur le lieu distant entraine une amélioration des performances (Pepper et Hightower, 1984). Si cette hypothèse est vraie, un système de télé-immersion avec une plus grande présence permettrait une meilleure collaboration des utilisateurs. Nous avons alors souhaité évaluer le sentiment de présence des utilisateurs de notre système de télé-immersion avec notre représentation 3D 360°. Dans la suite, nous nous plaçons du côté du visiteur, et nous chercherons à savoir si celui-ci préfère l'immersion dans notre représentation ou dans un simple flux omnidirectionnelle.

6.1.1 Présence

La présence est une notion largement étudiée en réalité virtuelle. Selon (Skarbez *et al.*, 2017), la présence est le réalisme perçu d'une expérience virtuelle ou médiatisée. La présence est alors une propriété subjective dépendant de la réalité perçue et le réalisme effectif est quant à lui une propriété objective dépendant des caractéristiques de l'expérience. La présence est intuitivement résumée à l'expérience subjective d'*être* dans une scène (Heeter, 1992). Aujourd'hui, le consensus admet que deux illusions orthogonales contribuent au sentiment de présence : l'illusion de place (*place illusion*, PI) et l'illusion de plausibilité (*plausibility illusion*, Psi) (Slater, 2009). L'illusion de place

correspond à l'illusion d'être dans un endroit malgré la certitude de ne pas y être, tandis que l'illusion de plausibilité correspond à l'illusion que ce qui se passe apparemment se passe réellement malgré la certitude que ce n'est pas le cas. L'illusion de place, ou présence spatiale, se réalise de manière plus ou moins convaincante selon un critère objectif du système appelé immersion. Dans (Slater, 2003), cette immersion est définie comme le niveau de fidélité sensorielle du système. Cela inclus par exemple la taille du champ de vision, la résolution de l'écran ou le taux de rafraîchissement (Bowman et McMahan, 2007). Le modèle de (Skarbez *et al.*, 2017) incorpore une troisième illusion pesant sur le sentiment de présence, l'illusion de présence sociale (*social presence illusion*). Nous suivons cependant une définition différente de celle proposée. En accord avec (Lombard et Ditton, 1997), nous posons la présence sociale comme l'illusion d'être avec un individu et d'interagir avec lui. Cet individu peut être un personnage virtuel ou une personne réelle et consciente. La coprésence est alors l'illusion d'être présent avec un individu, ensemble sur un même lieu. Présence sociale et coprésence sont souvent utilisées de manière interchangeable, mais le sentiment de coprésence repose sur le sentiment conjoint de présence spatiale et de présence sociale. La présence sociale peut varier en fonction de divers facteurs, comme la représentation visuelle des individus, les indices de profondeur (stéréoscopie et parallaxe de mouvement), ou même le retour haptique (Oh *et al.*, 2018).

Notre hypothèse est que le niveau de présence sur le lieu d'un visiteur dépend de la représentation de la scène. En particulier, nous supposons que le sentiment de présence sera plus grand dans notre représentation 3D 360° qu'avec une vidéo omnidirectionnelle en *skybox*. Ceci devrait se traduire par une observation d'une plus grande présence spatiale et d'une plus grande présence sociale avec notre représentation 3D 360°.

6.1.2 Présence dans un Environnement 360° et 3D

Pour vérifier si le niveau de présence varie entre une représentation omnidirectionnelle classique et notre représentation, nous avons passé en revue les différences entre une scène 360° (*3-DoF*) et une scène 3D (*6-DoF*) qui pourraient avoir une influence. Une caractéristique majeure de la caméra omnidirectionnelle est qu'elle permet de créer des scènes de manière plus photoréalistes que des scènes 3D modélisées par ordinateur. Or, avoir un rendu photoréaliste, c'est-à-dire un rendu qui correspond plus à ce qui pourrait être observé dans le monde réel, résulte objectivement en un niveau d'immersion supérieur. Il a donc été observé que le photoréalisme renforce effectivement l'illusion de lieu et donc le sentiment de présence (Zibrek *et al.*, 2019; Zibrek et McDonnell, 2019). L'augmentation du photoréalisme de la scène peut se traduire par l'ajout de détails dans la scène (Welch *et al.*, 1996), un plus grand nombre de polygones pour une géométrie plus fine (Hvass *et al.*, 2017), ou des textures plus réalistes (Lucaci *et al.*, 2022). Dans

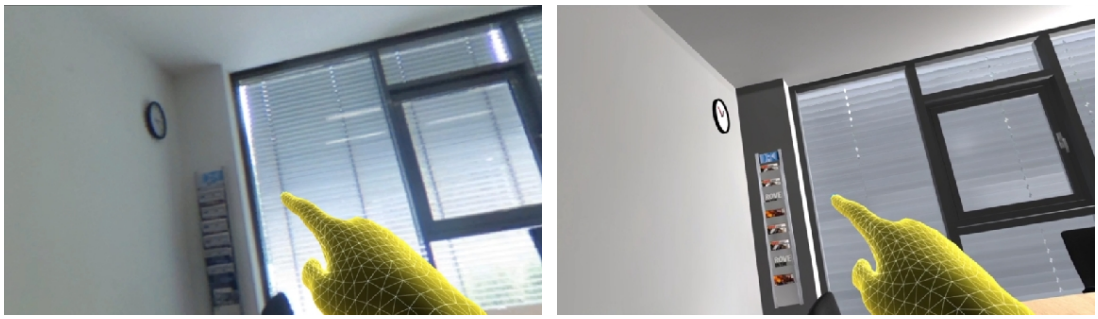


FIGURE 6.1 – Comparaison d'un même environnement sous forme d'image omnidirectionnelle et de modèle 3D (Schäfer *et al.*, 2021). Gauche : Image omnidirectionnelle du lieu. Droite : Réplique 3D reproduit par un artiste.

chacun de ces cas, une amélioration du sentiment de présence est observée. En revanche, même s'il est observable, cet effet paraît faible (Zibrek *et al.*, 2019) et semble marginal comparé à l'effet de l'interactivité de la scène (Welch *et al.*, 1996). Car c'est l'inconvénient de l'immersion dans une vidéo omnidirectionnelle avec une *skybox* : l'interaction est très limitée. La navigation en particulier est réduite à la seule orientation de la tête. Nous avons alors considéré une nouvelle définition de l'immersion proposée par (Slater, 2009). Dans cette version, l'immersion correspond à l'ensemble des actions valides supportées par le système. Une action valide est une action sensorimotrice qui permet de changer la perception de la scène ou une action effective qui permet de modifier l'état de la scène. Avec cette définition, l'immersion est réduite en l'absence d'interaction comme changer de point de vue ou déplacer des objets de la scène. Cette absence d'interaction, et notamment de navigation, impacte négativement le sentiment de présence. Il a donc été observé une diminution de la présence quand la parallaxe de mouvement est absente du système (Eftekharifar *et al.*, 2020; Barfield *et al.*, 1997) ou quand l'utilisateur ne peut pas choisir son point de vue (Lo et Lai, 2023; Clemente *et al.*, 2014). Ces observations permettent de conclure qu'il semble y avoir une rivalité entre le photoréalisme de la caméra omnidirectionnelle et le manque d'interactivité de l'image et la vidéo, rendant difficile de déterminer si le niveau de présence provoqué par cette représentation est supérieur à celui d'une représentation 3D.

Pour supporter notre hypothèse, nous avons étudié différentes expériences utilisateurs comparant directement la présence dans des scènes modélisées par images omnidirectionnelles en *skybox* avec celle dans des scènes 3D équivalentes. (Schäfer *et al.*, 2021) réalisent cette comparaison en modélisant en 3D des bureaux capturés par une caméra omnidirectionnelle. L'utilisateur est immergé dans les images omnidirectionnelles et dans des reproductions 3D pour comparer le sentiment de présence avec ces deux représentations (figure 6.1). Les auteurs insistent sur l'intérêt de l'image omnidirectionnelle pour avoir rapidement une scène du lieu, contrairement à la modélisation

3D par un artiste qui demande beaucoup de temps de création. Les résultats montrent un sentiment de présence équivalent entre l'image omnidirectionnelle et la reproduction en 3D. (Brivio *et al.*, 2021) arrivent à une même conclusion en comparant une scène extérieure capturée par une caméra omnidirectionnelle avec une autre similaire modélisée par ordinateur. Dans cette expérience, la trajectoire dans l'environnement est imposée et l'utilisateur ne peut que changer son orientation de la tête dans les deux représentations. On peut supposer que la différence dans le sentiment de présence induit par le photoréalisme de la caméra omnidirectionnelle est trop faible dans ces deux expériences pour être observable. (Nason *et al.*, 2020) comparent une vidéo omnidirectionnelle dans une boutique avec une reproduction 3D simpliste pour le traitement de l'anxiété sociale. Dans la scène 3D, l'utilisateur se déplace librement dans la scène 3D, tandis que dans la vidéo omnidirectionnelle l'utilisateur suit la trajectoire de la caméra. Les résultats montrent que la vidéo omnidirectionnelle suscite un sentiment de présence plus élevé, mais que l'impossibilité de naviguer librement peut créer plus d'anxiété. Dans cette expérience, l'effet du photoréalisme est observable car le niveau de réalisme de la scène 3D est bien inférieur à celui de la vidéo omnidirectionnelle. (Higuera-Trujillo *et al.*, 2017) comparent aussi le sentiment de présence entre une image omnidirectionnelle d'une boutique et une reproduction 3D photoréaliste du lieu où les utilisateurs se déplacent librement. Les résultats suggèrent que l'image tend à obtenir les meilleurs résultats sur le plan psychologique tandis que la reproduction 3D obtient les meilleurs résultats sur le plan physiologiques, rendant impossible de conclure sur la supériorité d'une représentation sur l'autre en termes de présence. (Boukhris *et al.*, 2017) comparent un ensemble d'images omnidirectionnelles d'une grotte avec une reconstruction par scanners 3D. L'utilisateur navigue dans la scène grâce à des points de téléportations prédéfinis qui sont les mêmes dans les deux représentations. Les résultats montrent une plus grande présence avec la reconstruction 3D, expliqués par l'absence de stéréoscopie et de parallaxe de mouvement sur une image omnidirectionnelle. La reconstruction 3D étant photoréaliste, l'image omnidirectionnelle ne profite plus de son avantage, son manque d'interactivité résulte en une présence plus faible. La comparaison entre 360° et 3D a été réalisée par (Ritter et Chambers, 2021) dans un contexte de formation. Dans cette expérience, l'arrière-plan a été modélisé à l'aide d'objets 3D dans un cas et avec une image omnidirectionnelle dans l'autre. Les objets de premier-plan sont dans les deux conditions des objets 3D. Les résultats montrent que le sentiment de présence mesuré est légèrement en faveur de l'environnement 3D, mais l'environnement 360° permet une plus grande efficacité dans la compréhension des concepts. Une expérience qui ressemble à l'évaluation que nous souhaitons conduire est celle de (Dupont de Dinechin et Paljic, 2018). L'étude compare le visionnage classique d'une image omnidirectionnelle avec une méthode de projection sur un proxy 3D, similaire à notre approche. Les

utilisateurs sont testés avec différents modes de visionnage (assis et debout) et avec des proxys de différentes précisions. L'expérience conclut que les utilisateurs préfèrent la reconstruction 3D dans la condition debout où ils peuvent se déplacer, mais ceux-ci préfèrent le visionnage sans relief dans la condition assise à cause des artéfacts. L'expérience ne donne cependant aucun résultat vis-à-vis de la présence. Enfin, l'expérience de (Arshad *et al.*, 2021) est aussi intéressante car elle compare le visionnage d'une vidéo omnidirectionnelle de manière classique avec le visionnage en 3D (via un proxy). Ceux-ci montrent que l'utilisation de la 3D avec la vidéo omnidirectionnelle réduit la sensation de cybermaladie. Or, si la cybermaladie et la présence sont bien inversement corrélées (Weech *et al.*, 2019), on peut supposer que ce mode de visionnage augmente la présence.

Pour conclure, certaines expériences semblent confirmer l'hypothèse que l'interactivité, en particulier la navigation, est plus importante que le photoréalisme pour le sentiment de présence. Même si des différences sont observées, la supériorité de la scène 3D sur la scène 360° pour le sentiment de présence n'est pas clairement démontrée.

6.1.3 Présence et Avatar

Dans un système de télé-immersion, le choix de l'incarnation n'est pas neutre pour les utilisateurs. Les différentes manières de représenter un utilisateur en réalité virtuelle influencent également la manière dont les autres le percevront. Comme relevé plus haut, la présence sociale fait partie des phénomènes affectés par l'apparence de l'incarnation. De nombreux auteurs ont alors entrepris d'identifier les éléments de l'incarnation en avatar qui impactent la présence sociale (Herrera *et al.*, 2020; Yoon *et al.*, 2019; Zibrek *et al.*, 2019; Zibrek et McDonnell, 2019; Smith et Neff, 2018; Heidicker *et al.*, 2017). Une partie des travaux se penche notamment sur la comparaison des différentes incarnations d'une même personne, et en particulier sur les avatars 3D de personnage. Ces derniers sont les avatars 3D, personnalisés ou génériques, qui sont connus à l'avance et animés par cinématique inverse en fonction des mouvements de l'utilisateur qui l'incarne. (Yu *et al.*, 2021) comparent cet avatar 3D de personnage avec un avatar 3D reconstruit en direct sous forme de nuage de points. Avant l'expérience, le participant personnalise l'apparence de son avatar à l'aide d'un logiciel de création d'avatars pour que son personnage lui ressemble. D'après les résultats, l'avatar en nuage de points présente un niveau de coprésence supérieur à l'avatar de personnage. (Jo *et al.*, 2017) ont aussi monté une comparaison similaire, entre un avatar 3D de personnage et un avatar 3D en maillage reconstruit en direct. Leurs résultats vont dans le sens inverse de ceux de l'expérience précédente et trouvent un niveau de coprésence plus élevé avec l'avatar de personnage pré-construit qu'avec l'avatar reconstruit en temps réel. Ces résultats sont justifiés par le phénomène de vallée de l'étrange (Mori *et al.*, 2012) qui rend l'avatar

FIGURE 6.2 – Comparaison de l'incarnation d'un même utilisateur sous forme d'avatar 3D et de *billboard* (Cho *et al.*, 2020). Gauche : Avatar volumétrique 3D. Droite : Avatar *billboard*.

reconstruit en temps réel moins crédible à cause des artéfacts.

Dans notre représentation 3D 360°, nous supposons que les interactions du système se concentrent sur un élément d'intérêt. Pour des raisons de performance, nous représentons les hôtes en mode dégradé. Inspiré par Tour Into Picture (Horry *et al.*, 1997), nous avons choisi d'utiliser des *billboards*, qui permettent de conserver un niveau de photoréalisme sans reconstruction 3D volumétrique. Cependant, nous nous sommes tout de même intéressés à savoir si la représentation d'utilisateur avec des avatars *billboard* est un choix raisonnable du point de vue d'un visiteur. Nous voulons déterminer si la présence est affectée par des incarnations d'avatars *billboards*. La perception des *billboards* a été le sujet d'expériences avec une visualisation non-immersive (Fourquet *et al.*, 2007; Hamill *et al.*, 2005) et immersive (Livatino, 2007). À notre connaissance, peu d'études comparent la perception des objets en 3D volumétrique avec des équivalents en *billboard*. L'effet sur la présence de ces différentes représentations pour des avatars en réalité virtuelle est donc peu connu. Pour notre hypothèse, nous relevons deux expériences pertinentes. La première expérience pertinente est celle de (Cho *et al.*, 2020)qui vise à étudier la perception de trois différentes incarnations : un avatar 3D volumétrique (figure 6.2 gauche), un avatar *billboard* (figure 6.2 droite) et un avatar 3D volumétrique pré-scanné. Un participant immergé dans une scène interagit avec un acteur incarné avec l'une des représentations. Tandis que l'avatar 3D volumétrique et l'avatar *billboard* sont créés à la volée, l'avatar 3D volumétrique pré-scanné est réalisé à l'avance et animé en temps réel. L'expérience se base sur deux tâches distinctes : une première dans laquelle le participant se déplace dans la scène, voyant ainsi la représentation de l'acteur sous différents points de vue, et une seconde dans laquelle le participant reste fixe face à l'acteur. Les résultats montrent que, dans la condition où le participant reste fixe, la présence sociale est équivalente entre l'avatar 3D volumétrique et le *billboard*. Mais

dans la condition où le participant se déplace, la conclusion est que la présence sociale est plus grande avec l'avatar 3D volumétrique qu'avec le *billboard*, certains participants relèvant le manque de relief du *billboard*. Quant à la présence sociale de l'avatar 3D volumétrique pré-scanné, elle n'est pas supérieure à celle des autres représentations dans les deux conditions à cause du manque de réalisme visuel. L'autre expérience pertinente a été réalisée par (Debarba *et al.*, 2022). Dans celle-ci, les utilisateurs sont immergés dans une même scène, sous forme de vidéo omnidirectionnelle en *skybox*, de représentation 3D volumétrique et de représentation 3D hybride. Dans cette dernière représentation, la majorité de l'environnement est reconstruit en 3D volumétrique, mais les personnes de la scène sont modélisées avec des *billboards*. Cette représentation hybride est alors assez similaire à notre représentation 3D 360°. Les résultats de l'expérience indiquent que la préférence des utilisateurs se portent vers la représentation 3D hybride. Le sentiment de présence globale dans la représentation 3D hybride est équivalent celui de la représentation 3D volumétrique, et supérieur à celui de la représentation omnidirectionnelle en *skybox*. La capacité de se déplacer à travers les représentations 3D, permettant aux utilisateurs de se rapprocher des avatars, renforce le sentiment de présence. Les auteurs expliquent également que la qualité vidéo pour l'apparence des avatars favorise la représentation 3D hybride. Bien que les *billboards* manquent de relief et présentent des incohérences avec l'environnement 3D, ces artéfacts ont significativement moins d'impact que la possibilité de se déplacer ou la meilleure qualité visuelle des personnages.

De manière similaire aux études dirigées pour concevoir des méthodes de compression sans baisse de qualité perçue, nous souhaitons vérifier si l'incarnation simplifiée des utilisateurs en *billboards* ne dégrade pas la présence sociale. Notre expérience utilisateur vise à évaluer la présence de notre système sur deux points, la présence spatiale et la présence sociale. Nous avons alors utilisé une sous-représentation 3D 360° de celle présentée au chapitre 5. Dans la suite, nous avons uniquement conservé l'image omnidirectionnelle avec la carte de profondeur pour l'environnement et les *billboards* pour les personnes. Nous avons considéré seulement les personnes comme éléments de premier plan, le modèle 3D de l'objet d'intérêt n'est donc pas inclus. Bien qu'aucunes expériences n'aient directement comparé l'immersion dans une vidéo omnidirectionnelle avec notre représentation 3D 360°, aucune conclusions des expériences citées ne réfutent l'hypothèse que notre représentations uscite plus de présence qu'une *skybox*.

6.2 Évaluation de la Représentation 3D 360°

La question de la présence dirige le choix de la représentation à utiliser pour un environnement et pour des avatars. Afin d'améliorer la crédibilité de notre système et encourager son adoption, nous avons voulu avoir des retours sur la présence générée avec notre représentation 3D 360°. Nous avons aussi voulu avoir des retours sur des

mesures couramment utilisées pour évaluer des nouveaux systèmes. L'utilisabilité et la charge de travail perçue ont donc aussi été considérées pour l'évaluation.

Cette évaluation compare la modélisation d'une même scène avec deux conditions : 360 et 3D. La condition 360 est la modélisation de la scène sous forme d'images omnidirectionnelles visualisées avec une *skybox*. La condition 3D est la modélisation de la scène avec notre représentation 3D 360° simplifiée (à partir des mêmes images omnidirectionnelles). La première question de recherche de notre évaluation porte sur la relation entre la géométrie et la présence :

Q1 : Est-ce que le sentiment de présence est affecté par la condition ?

La seconde question est relative à l'utilisabilité de notre système de télé-immersion :

Q2 : Est-ce que l'utilisabilité est affectée par la condition ?

Enfin, la dernière question concerne la charge de travail perçue de notre système :

Q3 : Est-ce que la charge de travail est affectée par la condition ?

À ces questions, nous posons les hypothèses suivantes.

H1 : Les conditions 360 et 3D suscitent des niveaux de présence différents.

H2 : La présence est inférieure avec la condition 360.

Ces hypothèses s'appuient sur les études comparant la vidéo omnidirectionnelle avec des reconstructions 3D, qui observent une différence souvent en faveur de la 3D.

H3 : Les conditions 360 et 3D ont une utilisabilité équivalente.

H4 : Les conditions 360 et 3D ont une charge de travail équivalente.

Ces hypothèses se basent sur l'idée que l'ajout de la 3D ne rend pas le système plus complexe. Dans notre expérience, la variable indépendante est donc la modélisation de la scène qui peut prendre deux valeurs, 360 pour la scène sous forme de *skybox* et 3D pour la représentation 3D 360°. Les variables dépendantes sont la présence, l'utilisabilité et la charge de travail.

6.2.1 Protocole

La comparaison entre 360 et 3D est effectuée avec le visionnage d'une même vidéo omnidirectionnelle (scène dynamique) sous deux formes. Cette vidéo omnidirectionnelle de 3 minutes 20 secondes (résolution 3840×1920) est tournée dans la *VR ROOM* du LSI au CEA-List. La caméra est placée au centre de la pièce et deux personnes donnent oralement des explications sur le fonctionnement de la salle. Les explications sont en anglais pour être compris des participants non-francophones. La vidéo est destinée à être visualisée en réalité virtuelle, avec une *skybox* pour la condition 360, et avec notre reconstruction 3D 360° pour la condition 3D. La reconstruction 3D 360° est réalisée à l'avance à partir de cette vidéo.

Pour évaluer la présence des participants, nous avons utilisé le TPI (*Temple Presence Inventory*), un questionnaire développé dans le contexte de la téléprésence, décomposé en plusieurs catégories de questions pour évaluer la présence sous plusieurs aspects. Nous avons utilisé la version originale en anglais et conservé les questions des catégories *Spatial Presence* pour la présence spatiale, *Social Presence - Actor Within Medium* pour l'interaction parasociale de la présence sociale, *Social Presence - Passive Interpersonal* pour la présence sociale interpersonnelle passive, *Engagement* pour l'engagement et *Social Richness* pour la richesse sociale. Ces questions sont évaluées sur une échelle de Likert en 7 points pour choisir une valeur entre deux pôles extrêmes (*jamais - toujours...*). L'utilisabilité est évaluée avec le questionnaire SUS (*System Usability Scale*). Le SUS comporte 10 questions sur une échelle de Likert en 5 points pour donner son degré d'approbation (*pas du tout d'accord - tout à fait d'accord*). Enfin, la charge de travail est évaluée avec le NASA-TLX. Le NASA-TLX se compose de 6 questions sur une échelle de Likert en 7 points pour se prononcer aussi entre deux valeurs opposées. Un questionnaire préliminaire est aussi rempli par les participants avant l'expérience.

Pour cette expérience, 13 participants (3 F, 10 H) ont été recrutés, âgés en moyenne de 23.8 ans ($\sigma = 2.6$), principalement des étudiants dans le cadre de travaux pratiques sur la réalité virtuelle à Grenoble-INP en génie industriel, et quelques membres du LSI. Sur une échelle de Likert de 1 à 5 les participants ont une familiarité moyenne de 3.5 ($\sigma = 1.2$) avec la réalité virtuelle et de 2.5 ($\sigma = 1.5$) avec les vidéos 360°. L'expérience suit un plan intra-sujet où les participants sont soumis aux deux conditions 360 et 3D, désignées respectivement par A et B auprès d'eux, donnant un total de 26 observations. Afin de réduire l'effet d'ordre d'exposition aux conditions, les participants ont aléatoirement été divisés en deux groupes. Le premier groupe (6 participants) teste la condition 360 puis la condition 3D. Le second groupe (7 participants) teste la condition 3D puis la condition 360. Dans les deux conditions, les participants visualisent la scène avec un casque de réalité virtuelle en étant debout et en tenant les manettes. Pendant le visionnage avec la condition 360, la seule interaction disponible est l'orientation du champ de vision. Ce changement d'orientation est réalisé naturellement en orientant la tête vers une autre direction. Une interaction de *snap-turn* est aussi intégrée, commune en réalité virtuelle, pour décaler l'angle d'orientation vers la droite ou la gauche avec le joystick de la manette sans tourner la tête. Le visionnage avec la condition 3D propose les mêmes interactions que la condition 360, en ajoutant la possibilité de se déplacer librement pour voir la scène d'une autre position que celle de la caméra. Ce changement de point de vue peut être exécuté en se déplaçant physiquement (mouvement de la tête) ou avec à une métaphore de téléportation. Les participants sont libres de choisir leurs positions dans la scène et aucune instruction les incitant à se déplacer ne leur est donnée. Avant de regarder la vidéo dans chaque condition, une phase d'entraînement est réalisée. Dans

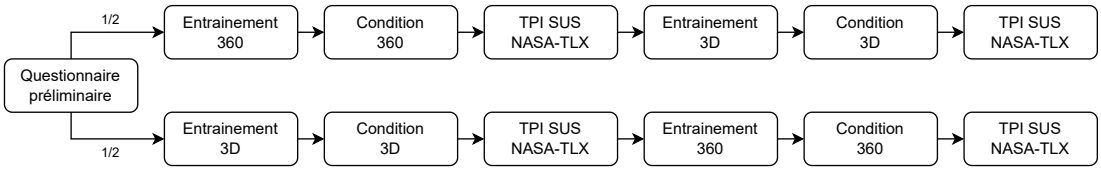


FIGURE 6.3 – Plan d'expérience de comparaison entre 360 et 3D.

cette phase, le participant est immergé dans un environnement statique, créé à partir d'une image omnidirectionnelle de la *VR ROOM* vide, afin de se familiariser avec les interactions. Pour la condition 360 la scène d'entraînement est une image omnidirectionnelle en *skybox*, et pour la condition 3D la scène est une reconstruction 3D 360° à partir de cette image (pas de *billboards* car le lieu est vide). La vidéo est lancée quand le participant déclare être prêt. Les questionnaires de présence, d'utilisabilité et de charge de travail sont renseignés après le visionnage de la vidéo dans chacune des conditions. Ce plan d'expérience est résumé figure 6.3. Le protocole, le formulaire de consentement et les questionnaires utilisés sont en annexe B.

6.2.2 Résultats

Pour l'analyse du TPI, nous avons calculé la réponse moyenne pour chacune des catégories : SP la moyenne de *Spatial Presence*, SPA la moyenne de *Social Presence - Actor Within Medium*, SPP la moyenne de *Social Presence - Passive Interpersonal*, EN la moyenne de *Engagement* et SR la moyenne de *Social Richness*. Nous avons aussi calculé la présence moyenne AP comme moyenne de toutes les réponses. Les résultats du TPI sont donnés figure 6.4. Les résultats du SUS sont calculés en sommant les réponses aux questions puis en normalisant de 0 à 100. Les résultats du NASA-TLX sont aussi calculés en ramenant la somme des réponses entre 0 et 100. Les résultats du SUS et du NASA-TLX sont donnés figure 6.5.

Analyse de la Présence

Il est difficile de connaître la distribution des résultats sur chacune des conditions car il n'y a pas d'informations du TPI sur la distribution théorique de la présence. De plus, la taille de l'échantillon ainsi que la visualisation des données qui ne semble pas suivre une distribution normale nous ont poussé à traiter les résultats du TPI comme des variables n'étant pas normalement distribuée. Nous supposons quand même a priori que les présences des différentes conditions suivent la même loi de distribution car issues de la même population. H1 suppose que les distributions de la présence sont décalées, et H2 que celle de la 3D est supérieure à celle de la 360. Nous avons alors analysé les résultats avec le test de Mann-Whitney U. L'hypothèse nulle est qu'il n'y a pas

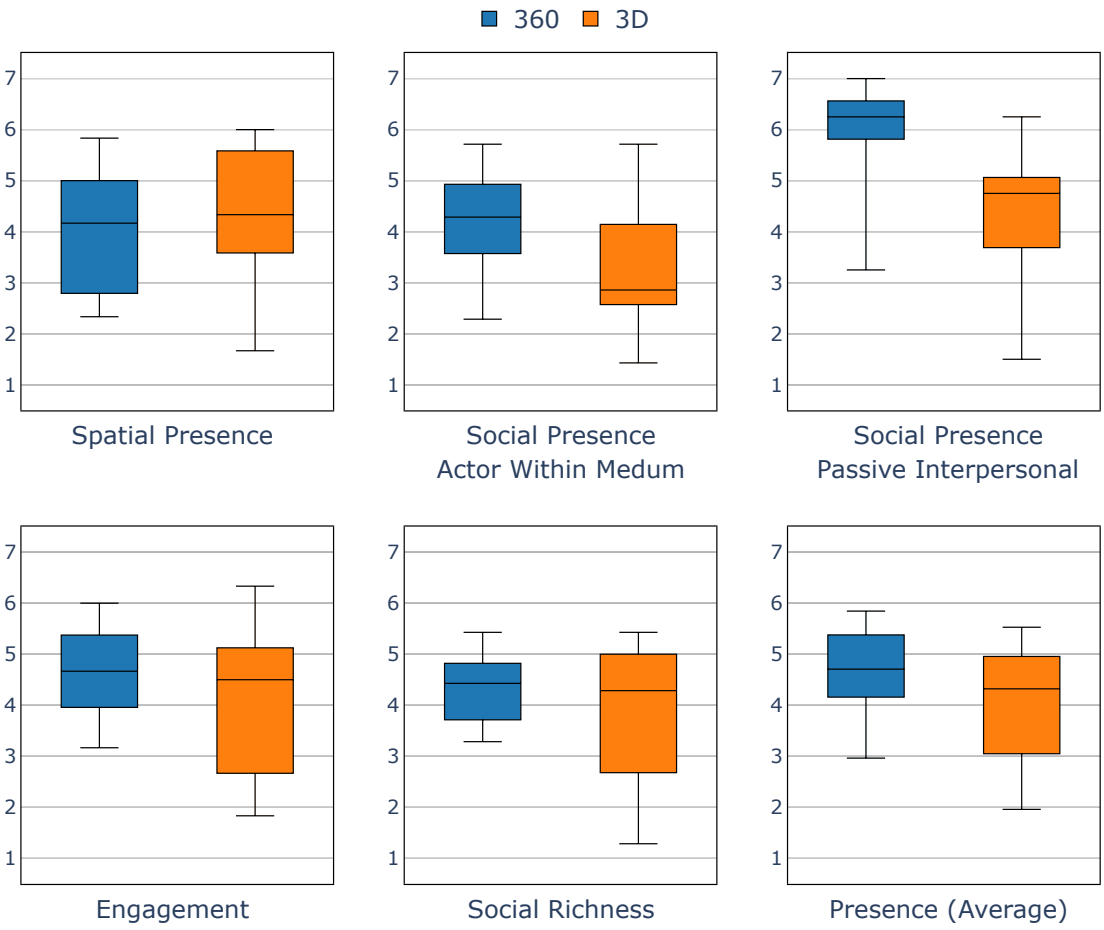


FIGURE 6.4 – Résultats du TPI.

de différences entre la présence médiane en condition 360 et la présence médiane en condition 3D. Ceci implique qu'il n'y a pas de différences sur AP entre 360 et 3D, mais aussi pas de différence sur chacune des catégories du TPI. Cette hypothèse à réfuter est la même pour valider H1 et H2. Pour H1, l'hypothèse alternative est bilatérale, pour H2 l'hypothèse alternative est unilatérale (360 inférieure à 3D). On pose le seuil de significativité à 0.05 avec une correction de Bonferroni. Comme 6 tests sont effectués, le seuil de significativité est ajusté à $\frac{0.05}{6} = 0.008$

On analyse d'abord s'il y a des différences significatives entre 360 et 3D. La valeur médiane de AP est de 4.7 pour 360 et 4.3 pour 3D. Cette différence n'est pas significative ($U = 60$, 0.218). La valeur médiane de SP est de 4.2 pour 360 et 4.3 pour 3D. Cette différence n'est pas significative ($U = 68.5$, $p = 0.426 > 0.008$). La valeur médiane de SPA est de 4.2 pour 360 et 2.9 pour 3D. Cette différence n'est pas significative ($U = 48$, $p = 0.064 > 0.008$). La valeur médiane de SPP est de 6.2 pour 360 et 4.8 pour 3D. Cette différence est significative ($U = 26$, $p = 0.003 < 0.008$). La valeur médiane de

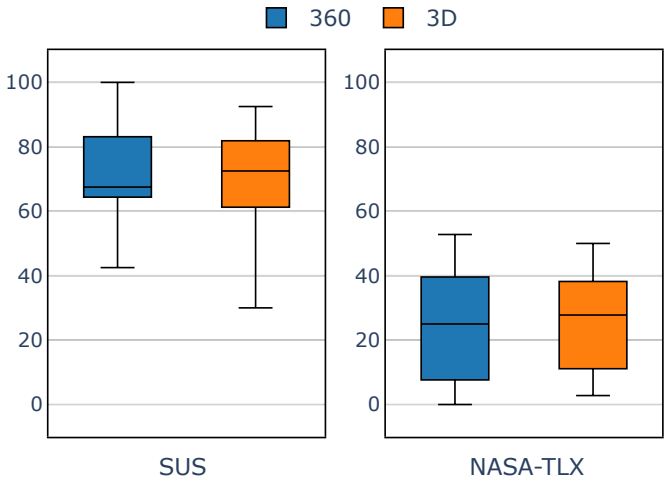


FIGURE 6.5 – Résultats du SUS et du NASA-TLX.

EN est de 4.7 pour 360 et 4.5 pour 3D. Cette différence n'est pas significative ($U = 69$, $p = 0.441 > 0.008$). La valeur médiane de SR est de 4.4 pour 360 et 4.3 pour 3D. Cette différence n'est pas significative ($U = 69.5$, $p = 0.456 > 0.008$). On a une différence significative pour SPP, l'hypothèse nulle est réfutée et H1 est validée.

Puisque que la seule différence significative entre 360 et 3D porte sur SPP, l'analyse d'infériorité est menée uniquement sur SPP. La valeur médiane de SPP étant plus grande pour 360 que pour 3D, on obtient une infériorité non-significative ($p = 0.999 > 0.05$). L'hypothèse alternative d'infériorité ne peut être vérifiée, H2 est rejetée.

Analyse de l'Utilisabilité et la Charge de Travail

Comme pour la présence, il n'y a pas de connaissances préalables sur la distribution des résultats, nos échantillons sont de taille réduite, et la visualisation ne suggère pas une distribution normale. Nous avons alors traité les résultats comme ne suivant pas une distribution normale. Nous supposons toujours que les distributions des deux conditions suivent la même loi, mais nous montrons ici qu'elles ont bien les mêmes paramètres (sans décalage). Pour vérifier H3, nous avons utilisé l'approche TOST. L'approche consiste en la réfutation de deux hypothèses nulles unilatérales pour affirmer que la différence de l'utilisabilité entre les deux conditions est bornée par une valeur Δ. L'idée est de supposer une première hypothèse nulle telle que la distribution des scores de 360 et la distribution des scores de 3D moins la valeur Δ sont égales (la médiane de 360 est égale à la médiane de 3D moins Δ). En réfutant cette première hypothèse nulle avec une hypothèse alternative supérieure, on montre que la distribution de 360 est plus grande que la distribution de 3D moins Δ (la médiane de 360 est supérieure à la médiane de 3D moins Δ). De manière symétrique, on pose une seconde hypothèse nulle telle que

la distribution de 360 et la distribution de 3D plus Δ sont égales. En réfutant cette seconde hypothèse nulle avec une hypothèse alternative inférieure, on montre que la distribution de 360 est plus petite que la distribution de 3D plus Δ. Avec ces deux réfutations, on montre que la distribution 360 est bornée par la distribution 3D moins Δ et 3D plus Δ : la différence entre les distributions 360 et 3D est donc de Δ. Or, si Δ est considéré comme suffisamment faible, alors les distributions sont considérées comme égales. La première hypothèse nulle H_0^- est donc que la distribution de l'utilisabilité de 360 est égale à celle de 3D-Δ, avec comme hypothèse alternative que la distribution de l'utilisabilité est plus grande avec 360 qu'avec 3D-Δ. La seconde hypothèse nulle H_0^+ est que la distribution de l'utilisabilité est de 360 est égale à celle de 3D+Δ, avec comme hypothèse alternative que la distribution de l'utilisabilité est plus petite avec 360 qu'avec 3D+Δ. Le choix de Δ est important pour considérer deux valeurs d'utilisabilité comme équivalente. Les scores du SUS standardisé (de 0 à 100) sont souvent interprétés en 5 catégories. On peut considérer que deux scores d'utilisabilité sont les mêmes s'ils sont dans la même catégorie. Avec une taille moyenne de catégorie de 20, deux scores s_1 et s_2 sont garantis d'appartenir à des catégories différentes si $|s_1 - s_2| > 20$. En dessous de cette valeur, nous considérons les scores équivalents car pouvant être dans la même catégorie. Δ est alors fixé à 20. Nous avons examiné H_0^- et H_0^+ avec le test Mann-Whitney U. Le seuil de significativité, fixé à 0.05 avec correction de Bonferroni pour 2 hypothèses, est à $\frac{0.05}{2} = 0.025$. Les utilisabilités médianes de 360 et 3D sont respectivement 67.5 et 72.5. La supériorité de 360 sur 3D-Δ est significative ($U = 25.5$, $p = 0.001 < 0.025$) et l'infériorité de 360 sur 3D+Δ est aussi significative ($U = 41.5$, $p = 0.015 < 0.025$). H_0^- et H_0^+ sont rejetées, H3 est donc validée.

Les mêmes suppositions et la même analyse sont utilisées pour vérifier H4. H_0^- pose que la distribution de la charge de travail de 360 est égale à celle de 3D+Δ et H_0^+ pose que la distribution de la charge de travail de 360 est égale à celle de 3D+Δ. Les hypothèses alternatives respectives sont que la distribution de la charge de travail est plus grande avec 360 qu'avec 3D-Δ, et que la distribution de la charge de travail est plus petite avec 360 qu'avec 3D+Δ. Le NASA-TLX étant aussi interprété en 5 catégories, la valeur de Δ est fixée 20. Les charges de travail médianes de 360 et 3D sont respectivement 25.5 et 27.8. Le test Mann-Whitney U montre une supériorité significative de 360 sur 3D-Δ ($U = 39$, $p = 0.01 < 0.025$) et une infériorité significative de 360 sur 3D+Δ ($U = 32$, $p = 0.004 < 0.025$). H_0^- et H_0^+ sont rejetées, H4 est donc validée.

6.2.3 Discussion

Validation de H1 et Rejet de H2

L'analyse de nos résultats nous permet de valider H1 mais pas H2. H1 est validée car une différence de présence sociale interpersonnelle passive est observée entre les conditions. La présence moyenne ainsi que les autres catégories ne montrent pas de différences significatives. En particulier, il est surprenant qu'une différence de présence spatiale ne soit pas observée. Une explication est que les différents participants n'ont pas tous agi de la même manière dans la condition 3D. Bien que libre de bouger dans la scène, certains participants sont restés à la position de la caméra, tandis que d'autres se sont largement déplacés tout le long de la vidéo. Ceux se déplaçant ont pu alors expérimenter une présence spatiale plus grande alors que ceux statiques n'ont pas ressenti de différence. Le problème peut résider dans le fait que notre expérience se base sur le visionnage d'une vidéo, le participant est plus passif qu'avec un système de télé-immersion où il interagit avec un utilisateur distant. En comparant les conditions 360 et 3D dans notre système de télé-immersion, la différence pourrait être accentuée sur des questions en rapport avec l'interaction parasociale ou la richesse sociale.

Comme l'infériorité de la condition 360 sur la condition 3D n'a pu être observée sur la présence moyenne ou les autres catégories, H2 n'est pas validée. La présence sociale interpersonnelle passive, seule différence observable, n'est pas significativement inférieur en 360 qu'en 3D, et celle-ci tend donc à être plus grande en 360 qu'en 3D. Une explication est que les questions sur la présence sociale interpersonnelle portent sur la perception des personnes sur site. Une explication est qu'après avoir terminé l'expérience, certains participants ont confié avoir préféré la condition 360 car elle permet d'avoir la meilleure vue de la scène. Dans cette condition, la perception d'une personne est meilleure car les participants restent face à elle, contrairement à la condition 3D où le *billboard* d'une personne présente des défauts selon la position du participant dans la scène.

Si les points faibles de notre représentation 3D 360° sont les avatars en *billboards* qui dégradent la présence, il est envisageable d'explorer différentes stratégies. L'expérience se basant sur le visionnage d'une vidéo, sans interactions avec la personne filmée, les avantages de la représentation 3D 360° ne sont pas évidents pour le participant. Une variante intéressante de l'expérience serait alors de comparer le système de télé-immersion complet, et non juste la représentation utilisée pour modéliser la scène. En mettant en avant l'interaction avec un utilisateur distant, le participant est moins passif qu'avec une vidéo. La possibilité avec la condition 3D de se déplacer en direct sur le lieu supplantera les inconvénients du *billboard* sur la présence. Une autre approche serait d'utiliser une représentation alternative pour les avatars dans notre représentation 3D 360°, par exemple un *billboard* articulé (Deng *et al.*, 2023; Germann *et al.*, 2010). Avec ce type de *billboard*, la vidéo de l'utilisateur n'est plus projetée sur un plan, mais sur une

surface 2D avec du relief, dans l'esprit du proxy. Si le relief épouse la pose de la personne filmée, avec des techniques d'estimation de pose ou de profondeur, l'avatar 3D est plus crédible, l'inconvénient restant que l'avatar 3D n'est toujours pas volumétrique. Cette représentation n'a pas été choisie pour notre représentation 3D 360° au chapitre 5 car nous avons supposé que le relief de l'avatar avait peu d'influence sur la présence. Mais à la vue de nos résultats, on peut faire l'hypothèse qu'en replaçant les *billboards* classiques par des *billboards* articulés, on influence positivement la présence sociale. Comme noté par (Debarba *et al.*, 2022), le manque de relief des *billboards* et ses incohérences avec l'environnement 3D ont pu être observés par les utilisateurs (même s'ils ont une influence mineure) et les *billboards* articulés résolvent précisément ces problèmes. Une autre modification pertinente serait de ne plus tester la représentation en *billboards* pour les personnes, mais pour les objets ambiants de la scène, c'est-à-dire ceux qui ne vont pas capturer l'attention d'un utilisateur. La cécité d'inattention (Cater *et al.*, 2002), c'est-à-dire la non-perception d'un phénomène par manque d'attention, peut jouer en faveur du *billboard*. Si l'utilisateur ne perçoit pas la différence entre une représentation volumétrique et un *billboard* pour des objets ambiants, le niveau de présence devrait être le même. Une future expérience serait alors de comparer des scènes avec des objets ambiants en 3D volumétrique et avec ces mêmes objets en *billboard*.

Pour conclure, bien que cette expérience ne montre pas que l'immersion dans notre représentation 3D 360° suscite un plus grand sentiment de présence que l'immersion dans une vidéo omnidirectionnelle classique, l'infériorité n'est pas non plus montrée. Notre représentation 3D 360° n'est pas inférieure en termes de présence à une image omnidirectionnelle en *skybox* et le temps de préparation de la scène est bien plus rapide qu'avec les méthodes de reconstruction 3D conventionnelles d'un lieu, par scan LiDAR ou reproduction par un artiste. Une autre expérience avec plus de participants conduira à plus de confiance dans les résultats obtenus.

Validation de H3 et H4

Nos résultats valident H3 et H4. L'utilisabilité du système est bien la même dans les conditions 360 et 3D. Toutefois, cela n'est pas totalement inattendu car le même dispositif de réalité virtuelle avec une interface similaire est utilisé dans les deux conditions. Les réponses du SUS sont influencées par la condition (3D ou 360) et par le dispositif en lui-même. Si l'influence du contenu visualisé est faible par rapport au dispositif, la variation d'utilisabilité entre les conditions peut ne pas être observée. Il est alors difficile de conclure définitivement que l'équivalence est due à la condition en elle-même plutôt que l'utilisation répétée du même dispositif. Les résultats montrent tout de même que l'ajout de la 3D ne compromet pas la simplicité d'utilisation perçue du dispositif de réalité virtuelle.

Les résultats pour la charge de travail perçue doivent aussi être nuancés. L'ajout de 3D, avec la nouvelle interaction de navigation, ne semble pas affecter les utilisateurs dans leurs charges de travail. Mais la charge de travail peut aussi ne pas avoir été affectée aussi car la tâche est simple dans les deux conditions. Le visionnage d'une vidéo laissant les utilisateurs passifs, la charge de travail de la tâche est globalement faible au point qu'une potentielle augmentation à cause la 3D pourrait ne pas être visible. Une future expérience, pour renforcer nos conclusions, devra comparer les conditions 360 et 3D, mais aussi analyser l'influence du dispositif de réalité virtuelle en utilisant plusieurs d'entre eux (casques, CAVE, ou même dispositif non-immersif). On peut tout de même conclure là aussi que la 3D n'augmente pas la charge de travail au point que le visionnage de la vidéo paraisse plus difficile.

6.3 Conclusion

Dans ce chapitre, nous avons évalué la représentation 3D 360° sur laquelle se base notre système de télé-immersion achevé au chapitre précédent. Une expérience utilisateur a été menée pour récolter le ressenti subjectif de cette représentation. Le niveau de présence étant une information cruciale pour comparer des scènes de réalité virtuelle, ce critère a été retenu pour évaluer notre représentation 3D 360°. Nous avons également examiné l'utilisabilité et la charge de travail associées à celle-ci. Nous avons souhaité comparer notre système avec l'approche de télé-immersion actuelle basée sur une caméra omnidirectionnelle. Aujourd'hui, l'approche de télé-immersion 360° consiste à simplement immerger le visiteur dans le flux vidéo omnidirectionnelle avec une *skybox*. L'interaction étant limitée avec cette approche, nous avons supposé que notre système susciterait plus de présence chez l'utilisateur. Nous avons alors comparé la présence lors du visionnage d'une même vidéo omnidirectionnelle, avec une *skybox* et avec notre représentation 3D 360°. Nos résultats ne permettent pas de trancher en faveur de notre représentation 3D 360°. Ceci peut s'expliquer notamment par l'absence d'interaction avec l'hôte lors du visionnage d'une vidéo, l'utilisateur ne percevant pas l'intérêt de la navigation libre. Ces résultats restent tout de même pertinents pour notre système car les artéfacts (surfaces fantôme, manque de relief du *billboard*...) ne semblent pas avoir d'effets significatifs sur la plupart des facteurs de présence testés, seulement la présence sociale interpersonnelle passive semble affectée. L'intérêt de notre représentation 3D 360° est alors que le lieu commun est rapidement modélisé en 3D pour un sentiment de présence des visiteurs qui n'est pas inférieur à une *skybox*. Une nouvelle hypothèse à tester pour une future expérience serait alors que la présence avec notre système est équivalente à celle de l'approche 360°. Pour l'utilisabilité et la charge de travail, nos résultats montrent qu'elle est équivalente entre les deux systèmes. L'ajout de la 3D ne rend pas le système de télé-immersion plus complexe, malgré une plus grande immersion

et interactivité.

Chapitre 7

Conclusion

Au début de ce manuscrit, nous avons dessiné les contours d'un système de télé-immersion idéal à atteindre. Ce système idéal est une solution prometteuse pour supporter des cours en ligne où des étudiants à distance participent activement aux leçons dispensées par un professeur comme s'ils étaient physiquement dans la salle de classe. Le cadre théorique que nous avons élaboré met en évidence l'intérêt de la caméra 360° et des casques de réalité étendue pour une télé-immersion 3D nomade avec une immersion totale. Nous avons successivement développé des systèmes de télé-immersion 3D 360° pour se rapprocher de notre système idéal. La version définitive de notre système a été le sujet d'une expérience utilisateur pour faire ressortir ses avantages par rapport à la diffusion de vidéo en télé-immersion 360° traditionnelle. Dans ce chapitre, nous présentons la synthèse de nos travaux de recherches, les contributions à l'état des connaissances actuelles et les réponses à nos questions de recherche. Des perspectives sont également ouvertes sur les domaines théoriques et techniques nécessitant une exploration plus profonde.

7.1 Synthèse des Travaux de Recherche

La télé-immersion se penche sur l'étude des systèmes qui rassemblent des utilisateurs géographiquement éloignés comme s'ils étaient tous physiquement sur un lieu commun. L'ensemble de ces systèmes forme une grande diversité, allant des logiciels de visioconférence aux robots humanoïde en passant par la reconstruction 3D de personnes ou d'environnements. Au chapitre 1, nous avons proposé d'enrichir la liste des systèmes de télé-immersion avec une contribution originale basée sur une caméra omnidirectionnelle et des casques de réalité étendue. Nous avons alors posé le système idéal que nous désirions atteindre et les questions de recherche relatives à son développement. Cette section résume les contributions apportées à chaque chapitre et répond aux questions de recherche.

7.1.1 Contributions

La réalisation de cette thèse a mené à quatre contributions : un nouveau cadre théorique de la télé-immersion, un système de télé-immersion 360° dans des environnements uniquement à partir d'images omnidirectionnelles, un système de télé-immersion 3D 360° dans des lieux statiques capturés par scans 3D, et un système de télé-immersion 3D 360° sur des lieux dynamiques, proche du système idéal.

Théorie de la Télé-Immersion

La télé-immersion a évolué organiquement en réunissant des chercheurs de diverses disciplines pour travailler sur des projets de collaboration à distance. Cette diversité d'horizons fait que la télé-immersion a émergé sans bénéficier d'un cadre théorique commun. Nous avons proposé au chapitre 2 un nouveau cadre théorique de la télé-immersion. Cet cadre se base sur les notions de lieu et de scène déjà définis. Un lieu est un emplacement sur lequel sont présents des utilisateurs. Ils y sont présents car ils sont physiquement sur le lieu ou parce qu'ils y sont virtuellement transportés. Le lieu où sont rassemblés les utilisateurs est le lieu commun. La scène est la représentation des éléments d'intérêt d'un lieu. Ces éléments d'intérêt peuvent être des utilisateurs ou un objet sur lequel ils doivent collaborer. La scène permet à un utilisateur d'amener un autre utilisateur distant auprès de lui sur son lieu physique, ou d'être immergé sur un lieu distant. La télé-immersion repose alors sur l'extraction de scène pour créer la scène à partir du lieu et sur l'inclusion de scène pour présenter la scène à un utilisateur.

Notre théorie insiste sur le fait que les utilisateurs n'ont pas un rôle équitable dans la télé-immersion. Certains se sentent présents sur leurs lieux physiques et accueillent des utilisateurs distants, d'autres se sentent présents sur un autre lieu qui n'est pas leurs lieux physiques. Ces deux rôles bien distincts sont appelés respectivement hôte et visiteur. On trouve alors trois types de télé-immersion : hôte-visiteur impliquant des hôtes et des visiteurs, hôte-hôte impliquant uniquement des hôtes et visiteur-visiteur impliquant uniquement des visiteurs. Cette différence de rôle permet de donner un sens aux notions de symétrie et asymétrie, souvent mises en avant dans les systèmes de télé-immersion.

Un autre apport dans notre théorie est de considérer qu'une scène est uniquement composée de deux types d'éléments : objet et environnement. Un objet est une représentation individuelle d'un élément particulier du lieu, comme la reconstruction 3D d'une personne ou d'un objet d'intérêt. Un objet a tendance à représenter un élément dans le champ proche d'un utilisateur avec lequel il va pouvoir interagir, comme tourner autour ou le manipuler. Un environnement est une représentation globale du lieu, comme une image omnidirectionnelle. Un environnement a plutôt tendance à représenter le champ lointain d'un utilisateur avec lequel il n'y a pas forcément d'interactions. Cette

distinction s'appuie sur une différence dans les dispositifs d'acquisition utilisés pour l'extraction de scène On remarque qu'ils sont qualifiés de *outside-in* pour capturer un élément précis, ou *inside-out* pour une vue globale. Nous posons alors que les dispositifs *outside-in* capturent des objets et *inside-out* des environnements. Une conséquence de cette distinction est que le lieu commun est uniquement déterminé par l'environnement et que les objets n'ont aucun effet. En effet, si un utilisateur est immergé dans une image omnidirectionnelle d'un premier lieu, on peut ajouter autant d'objets que possible d'un second lieu, l'utilisateur continuera de se sentir sur le premier lieu.

Enfin, pour l'inclusion de scène, nous nous sommes intéressés aux dispositifs qui permettent d'immerger un utilisateur dans une scène. Ceux-ci sont traditionnellement répartis entre immersion matérielle et immersion virtuelle. Les premiers concernent la robotique et les seconds la réalité étendue. Cependant, cette classification ne représente pas la diversité des incarnations, c'est-à-dire des représentations utilisées pour modéliser un utilisateur d'un lieu distant. Dans notre cadre théorique, quatre familles d'incarnation sont identifiées : vidéo, hybride, robot et avatar. Chacune est liée à des technologies de télé-immersion différentes. La vidéo est propre à la visioconférence avec ses variantes. La représentation hybride se caractérise principalement par l'utilisation de robots de téléprésence où l'utilisateur distant est affiché sous forme de vidéo avec un écran monté sur une base mobile. Le robot englobe les technologies de télé-opération collaborative et notamment les robots anthropomorphiques. Enfin, l'avatar consiste en les représentations humanoïdes 3D visualisées en réalité étendue, aussi bien en réalité virtuelle et en réalité augmentée avec écrans, casques, ou par projection.

Avec notre formalisme, le système de télé-immersion développé est asymétrique car il réunit dans une salle de classe comme lieu commun, un professeur qui joue le rôle d'hôte, et des étudiants qui jouent le rôle de visiteurs. L'extraction est effectuée avec une caméra omnidirectionnelle pour capturer l'environnement de la salle de classe. L'inclusion est réalisée avec des dispositifs de réalité étendue afin d'immerger les visiteurs dans la salle de classe et de les incarner auprès des hôtes en réalité mixte.

Télé-Immersion 360° dans un Environnement

Au chapitre 3, nous nous sommes concentrés spécifiquement sur la télé-immersion basée 360°. Nous avons passé en revue les différents types de caméras omnidirectionnelles, les données qu'elles capturent et comment immerger un utilisateur dans ces données. Nous avons pu en retirer des propriétés de l'extraction de scène basée sur une caméra omnidirectionnelle. La première est topologique, une caméra omnidirectionnelle capture la surface d'un convexe étoilé à l'intérieur du lieu. Le lieu est alors visible dans son intégralité si le lieu est un convexe étoilé et que la caméra est placée sur un centroïde. Cette propriété est très contraignante car si le lieu contient des objets, sa topologie ne peut pas

être en convexe étoilé. Néanmoins, si le lieu n'est qu'un environnement sans objets au premier plan, son approximation en convexe étoilé est raisonnable. Une autre propriété est que l'image omnidirectionnelle ne contient pas d'information géométrique. Les utilisateurs peuvent alors être immergés sur le lieu uniquement à la position de la caméra. L'immersion de plusieurs visiteurs incarnés en avatar est alors complexe car ils sont virtuellement tous à la même position. Notre première proposition de télé-immersion vise alors à immerger plusieurs visiteurs, incarnés en avatars, dans un environnement capturé par une caméra omnidirectionnelle, sans informations géométriques.

La règle normale pour la collaboration dans des scènes 3D est que si deux utilisateurs sont à des positions différentes, ils doivent voir la scène avec des points de vue différents. Or, cette règle est impossible à respecter avec une image omnidirectionnelle sans géométrie. L'idée de notre proposition est alors de briser cette règle et d'utiliser la même image omnidirectionnelle pour les points de vue de tous les utilisateurs. Notre approche donne l'illusion à chacun des utilisateurs que les autres voient la scène d'un autre point de vue. Pour cela, on ajoute simplement de manière cohérente un avatar pour tous les autres utilisateurs dans l'image omnidirectionnelle. La vue d'un utilisateur à une position différente de la sienne laisse penser qu'il est bien à une autre position et donc qu'il voit la scène d'un autre point de vue. Cette illusion semble fonctionner et l'approche est intéressante pour l'enseignement en ligne afin d'immerger plusieurs étudiants à distance dans une vidéo omnidirectionnelle du lieu commun. Cependant, des limites intrinsèques à la méthode nous ont poussé à ne pas poursuivre cette direction.

La première limite est que les utilisateurs peuvent objectivement se rendre compte qu'ils sont en réalité à la même position dans la scène et voient la même image omnidirectionnelle. Objectivement, l'illusion fonctionne parfaitement quand le lieu suit une topologie convexe. Dans ce cas, tous les éléments visibles par un utilisateur sont visibles par un autre. Mais si le lieu suit une topologie en convexe étoilé, des problèmes dans la cohérence des occultations peuvent apparaître. Dans ce cas, pour un utilisateur particulier, un autre utilisateur devrait ne pas voir certains éléments en raison du décalage apparent qu'il a avec lui. Or, celui-ci étant en réalité à la même position, ils voient de fait les mêmes éléments, laissant alors penser que l'autre utilisateur voit à travers les murs. La seconde limite concerne l'immersion et l'interaction. Le problème est l'impossibilité pour un élément de l'environnement d'occulter un utilisateur et l'impossibilité d'implanter une interaction comme un pointeur laser. Ceci est directement la conséquence du manque d'informations dans une image omnidirectionnelle, notamment d'informations géométriques. À noter que ces limites se présentent quand les éléments de l'image omnidirectionnelle sont interprétés comme proches. L'approche est donc relativement adaptée si les éléments sont lointains comme dans un environnement en extérieur.

Télé-Immersion 3D 360° dans un Lieu Statique

Au chapitre 4, nous avons intégré des informations géométriques pour une télé-immersion 3D basée 360°. L'incorporation d'informations 3D est motivée par le souhait de surmonter les limites du système précédent. Nous avons choisi de représenter les informations géométriques associées à une image omnidirectionnelle sous forme de carte de profondeur omnidirectionnelle. La carte de profondeur omnidirectionnelle est une image omnidirectionnelle en niveaux de gris où la valeur d'un pixel indique la distance entre ce point et la caméra. La carte de profondeur omnidirectionnelle permet de reconstruire un proxy en 3D, c'est-à-dire un volume sur lequel est projetée l'image omnidirectionnelle. En plaçant un utilisateur à l'intérieur du proxy, l'utilisateur est immergé dans une image omnidirectionnelle, mais l'immersion et l'interaction est améliorée par rapport au système précédent. Avec cette représentation, il est possible d'occulter un objet ajouté dans la scène avec un élément de l'image omnidirectionnelle, et de générer naturellement d'obtenir une parallaxe de mouvement. Un des enjeux est alors d'obtenir une carte de profondeur omnidirectionnelle à partir d'une caméra omnidirectionnelle. Nous avons vu que les méthodes d'estimation de carte de profondeur omnidirectionnelle reposent largement sur des réseaux de neurones prenant en entrée une ou plusieurs images. Les méthodes existantes sont sujettes à de nombreux artéfacts, comme des distorsions ou la perte de la géométrie des petits éléments, nuisant à la réalité du lieu pour un utilisateur. Nous avons alors proposé d'adapter une approche pour avoir une reconstruction 3D mitigeant ces artéfacts. Notre idée est d'exploiter une approche d'estimation sur des images perspectives car la base d'entraînement des réseaux de neurones perspectives est plus large et donc la reconstruction semble moins bruitée. Nous estimons une carte de profondeur d'une image omnidirectionnelle en la découpant en un ensemble d'images perspectives, en estimant la carte de profondeur de chaque image perspective individuellement, et en agrégeant l'ensemble des cartes de profondeur en une carte de profondeur omnidirectionnelle.

La carte de profondeur améliorant l'immersion et l'interaction dans une scène modélisée par une image omnidirectionnelle, nous avons proposé un nouveau système de télé-immersion 3D 360° améliorant celui du chapitre précédent. L'idée est d'immerger des utilisateurs dans un lieu à travers des images omnidirectionnelles avec des cartes de profondeur. Pour valider le concept en évitant les problèmes liés à l'estimation de la carte de profondeur, nous avons considéré le cas où la géométrie complète du lieu commun est connue à l'avance. Nous avons alors développé notre système pour une application de navigation dans des scans 3D de lieux sous forme de nuages de points. Pour ne pas traiter non plus les problématiques liées à la génération de la scène en temps réel, nous avons uniquement considéré les nuages de points statiques. Avec notre système, plusieurs visiteurs peuvent se retrouver sur un nuage de points stocké sur un serveur. Le

serveur envoie le rendu du nuage de point à la position d'un visiteur sous forme d'image omnidirectionnelle et de carte de profondeur. Un utilisateur voit les autres visiteurs incarnés en avatar dans l'image avec une gestion correcte des occultations grâce à la carte de profondeur. L'utilisateur se déplace aussi dans l'image omnidirectionnelle, avec pour limites que des artéfacts deviennent de plus en plus apparents lorsqu'il s'éloigne de la position initiale de l'image. Enfin, l'utilisateur navigue sur le nuage de points grâce à une métaphore de téléportation. À l'aide d'un pointeur laser, celui-ci sélectionne au sol une destination à laquelle il souhaite se téléporter. Le déclenchement de la téléportation lance une requête au serveur pour obtenir un nouveau rendu aux coordonnées sélectionnées. Nos évaluations mettent en évidence une utilisation intermittente de la bande passante avec notre méthode car une communication continue avec le serveur n'est nécessaire que pour un changement de position. Nos résultats montrent aussi que notre méthode utilise moins de données que les approches standard de télé-immersion dans un nuage de points.

Malheureusement, l'image omnidirectionnelle avec la carte de profondeur présente toujours deux limites fondamentales. La première limite est que la représentation 3D avec un proxy considère encore l'image comme une seule et même surface. L'image omnidirectionnelle avec la carte de profondeur reste donc une modélisation de l'environnement de la scène et ne convient pas pour modéliser les objets de la scène. Par conséquent, il demeure impossible d'interagir avec des éléments individuels de l'image comme s'ils étaient des objets. La seconde limite est qu'aucune solution n'a été trouvée pour estimer une carte de profondeur en temps réel avec une qualité appropriée pour l'immersion en réalité virtuelle. La carte de profondeur ne peut donc pas être utilisée conjointement avec la vidéo omnidirectionnelle pour la télé-immersion sur des lieux physiques dynamiques. L'image omnidirectionnelle avec carte de profondeur reste tout de même pertinente pour la télé-immersion dans des environnements simples.

Télé-Immersion 360° dans un Lieu Dynamique

À partir des conclusions sur les systèmes précédents, nous avons proposé au chapitre 5 une dernière version de notre système de télé-immersion 3D 360°. Ce système constitue une avancée par rapport au précédent sur deux points. La première avancée est que la représentation 3D proposée pour modéliser la scène ne contient plus seulement la représentation de l'environnement mais aussi les représentations des objets. La seconde avancée est que cette représentation 3D est obtenue en temps réel, permettant de télé-immerger des visiteurs sur des lieux physiques dynamiques. Avec notre système, un enseignant peut facilement amener des étudiants à distance dans la salle de classe en temps réel avec une caméra omnidirectionnelle qui diffuse la vidéo en direct. Les étudiants, immergés avec des casques de réalité virtuelle, sont libres de se déplacer dans

la salle et d'interagir avec l'enseignant, les étudiants sur site et ceux à distance. L'enseignant et les étudiants sur site voient, avec des casques de réalité mixte, ceux à distance dans la salle de classe incarnés en avatar. Ce système a été développé pour supporter des cours où l'enseignant donne des explications sur un objet d'intérêt comme une machine.

La version définitive de notre système repose sur une nouvelle représentation 3D 360°, inspirée des représentations stratifiées. Trois types d'éléments composent notre représentation 3D 360°. D'abord, une image omnidirectionnelle avec carte de profondeur pour modéliser un environnement statique en relief, en accord avec la version précédente. Un modèle 3D aligné sur la vidéo omnidirectionnelle pour modéliser l'objet d'intérêt. Avec cette modélisation, l'objet d'intérêt est traité comme un objet classique de réalité virtuelle. Enfin, des *billboards* pour représenter les hôtes, c'est-à-dire un plan 2D positionné dans l'espace 3D sur lequel est projeté le flux vidéo de la personne.

Différentes étapes sont suivies pour générer ces différentes composantes de la représentation 3D 360°. La première étape est de recaler le modèle 3D de l'objet d'intérêt sur une image omnidirectionnelle. Pour cela, nous résolvons un problème de correspondance entre des images du modèle 3D et l'image omnidirectionnelle du lieu. La résolution de ce problème nous donne la position du modèle 3D de l'objet d'intérêt dans l'espace pour qu'il soit superposé sur l'image omnidirectionnelle. Nous obtenons ensuite l'environnement grâce à une image omnidirectionnelle de l'arrière-plan du lieu. Pour cela, les éléments de premier plan sont détectés sur une image omnidirectionnelle du lieu puis effacés. En considérant que l'objet d'intérêt et les personnes sont les seuls éléments de premier plan, nous combinons le résultat du recalage du modèle 3D avec une détection d'humain pour obtenir un masque des éléments à effacer. Ces éléments de premier plan sont effacés à l'aide d'un outil d'*inpainting*. La dernière étape pour l'environnement est d'estimer la carte de profondeur grâce à notre approche précédemment mise au point. À noter que ces étapes étant trop lentes pour être exécutée en temps réel, elles ne sont réalisées qu'à l'initialisation sur la première image du flux vidéo.

Les étapes suivantes sont réalisées en direct pour chaque image du flux vidéo. Le modèle 3D n'étant aligné que sur la première image du flux vidéo, nous utilisons une ombre numérique pour animer le modèle 3D. En synchronisant la diffusion des données de l'ombre numérique avec la diffusion du flux omnidirectionnelle, les mouvements du modèle 3D sont synchrones avec l'objet d'intérêt sur la vidéo. Nous utilisons également l'image courante de la vidéo pour texturer le modèle 3D afin de rendre son apparence plus réaliste. Pour créer les incarnations des hôtes, nous extrayons d'abord une image perspective centrée pour chacune des personnes. Ces images perspectives sont obtenues grâce à une segmentation de la vidéo équirectangulaire suivie d'une projection gnomonique par personne. À partir de ces images, on ajoute un plan dans l'espace pour chaque

personne. Les différentes propriétés de la projection équirectangulaire permettent de déterminer la position dans l'espace et la taille du *billboard* d'une personne.

Nos évaluations techniques objectives montrent que notre système avec une résolution moyenne est bien temps réel, que le détourage des *billboard* sur l'image omnidirectionnelle est de même qualité qu'avec une résolution élevée, et que la bande passante du réseau est comparable à la diffusion de la vidéo omnidirectionnelle. Nous avons alors conduit au chapitre 6 une expérience utilisateur pour une évaluation subjective. L'idée est d'avoir des retours d'utilisateurs d'un système basé sur notre représentation 3D 360° sur des informations comme le sentiment de présence, l'utilisabilité et la charge de travail. Nous avons donc comparé une version simplifiée de notre représentation 3D 360° (seulement l'environnement et les personnes en *billboards*) avec une simple vidéo omnidirectionnelle dans une expérience intra-sujet. Pour cela, nous avons immergé en réalité virtuelle des participants dans deux versions de la même vidéo, une version en vidéo omnidirectionnelle traditionnelle (sans 3D) et une version avec notre représentation 3D 360°. Nos hypothèses étaient que notre représentation susciterait plus de présence chez les participants, mais que l'utilisabilité et la charge de travail seraient équivalentes dans les deux conditions. La première hypothèse était justifiée par des expériences qui montraient que la présence était plus grande dans des scènes en 3D que dans leurs équivalentes en images omnidirectionnelles. La seconde hypothèse repose sur l'intuition que l'ajout de géométrie n'augmente pas la complexité du système. L'analyse de nos résultats confirme l'hypothèse que l'utilisabilité et la charge de travail sont les mêmes dans les deux conditions. Malheureusement, nos résultats ne confirment pas l'hypothèse que notre représentation 3D 360° induit un plus grand sentiment de présence car une présence supérieure en notre faveur n'a pas pu être observée. Une explication est que la nature passive du visionnage d'une vidéo n'encourage pas les participants à ressentir une plus grande présence avec notre représentation. Malgré tout, nous pouvons conclure que notre représentation 3D 360° ne fait pas pire que la simple diffusion d'une vidéo omnidirectionnelle. Notre système est alors une bonne base pour des futurs systèmes de télé-immersion.

7.1.2 Réponses aux Questions de Recherches

Au chapitre 1, nous avons posé les questions de recherche suivante :

Q1 : Est-il possible de représenter les avatars de plusieurs utilisateurs avec une unique caméra 360° ?

Q2 : Est-il possible de se déplacer librement sur un site capturé uniquement avec une caméra 360° ?

Q3 : Est-ce que les utilisateurs préfèrent la vue 360° avec ou sans navigation libre ?

À ces questions de recherches, nous avions répondu positivement. Les conclusions tirées lors de nos contributions nous permettent de répondre plus précisément à ces questions.

Pour *Q1*, nous avons supposé qu'il existe bien un moyen d'être immergé à plusieurs dans une image omnidirectionnelle. Le développement de notre système de télé-immersion 360° dans un environnement nous montre que c'est bien possible. La difficulté réside dans le fait qu'une image omnidirectionnelle ne peut représenter qu'un seul point de vue à la fois, ce qui exclut la possibilité d'avoir deux personnes à des positions distinctes dans un lieu capturé par une caméra omnidirectionnelle. La majorité des chercheurs suggèrent des incarnations de visiteurs adaptées à cette contrainte en les représentant à l'aide de rectangles sur l'image omnidirectionnelle afin de souligner leurs champs de vision. Notre approche a été d'ajouter simplement des incarnations avatars à l'image omnidirectionnelle de manière cohérente pour les autres visiteurs dans l'image omnidirectionnelle pour donner l'illusion qu'ils sont ailleurs sur le lieu. Mais cette illusion est fragile et peut être brisée par un manque de cohérence dans les occultations entre les utilisateurs. De plus, cette approche n'est pas une base idéale pour un système de télé-immersion 3D, car les éléments de l'image omnidirectionnelle ne peuvent pas occulter les incarnations et le point de vue d'un visiteur doit strictement être celui de la caméra sinon le rendu présentera un effet de flottement des avatars dans l'environnement.

Pour *Q2*, nous avons supposé qu'il existait une manière d'ajouter des informations 3D à un flux omnidirectionnelle. Les systèmes de télé-immersion 3D 360° que nous avons développé, pour l'immersion dans un lieu statique et dynamique, permettent de répondre aussi positivement à la question. La carte de profondeur ou notre représentation 3D 360° sont bien des images omnidirectionnelles avec géométrie. En plus de permettre plusieurs avatars de visiteurs dans la scène sans les problèmes précédents, ces représentations permettent bien de se déplacer dans une image omnidirectionnelle.

Pour *Q3*, nous avons supposé que la liberté de navigation était plus apprécié par les utilisateurs que le fait de rester figés à la position de la caméra. Nous avons fait en particulier l'hypothèse que les utilisateurs se sentiraient plus présent dans le lieu avec notre représentation 3D 360° qu'avec une simple vidéo omnidirectionnelle. Nous pensions que la navigation libre influencerait positivement la présence. Notre expérience utilisateur laisse cette question en suspens. Nous ne sommes actuellement pas capables d'affirmer que notre représentation 3D suscite une plus grande présence, bien qu'elle offre en évidence plus d'immersion et d'interaction.

7.2 Perspectives

La télé-immersion est un ensemble de technologies dans la continuité de celles du numérique pour l'enseignement à distance. Le système de télé-immersion final que nous avons développé peut toutefois se généraliser à de nombreux cas d'usage. Dans notre scénario, les hôtes sont debout autour de la caméra omnidirectionnelle et les visiteurs se déplacent librement sur le lieu commun, ce qui est acceptable pour des applications d'enseignement à distance. Mais ce système est aussi utilisable pour tenir des réunions mixtes présentielles distancielles. Dans un tel scénario, le système pourrait rassembler des hôtes autour d'une table, la caméra placée au centre la table, avec des visiteurs qui assistent à la réunion comme s'ils étaient autour de la table. Le déplacement libre permet simplement d'assigner une place aux différents visiteurs, comme une chaise libre. Dans ce cas, le manque de relief du *billboard* ou son manque de cohérence avec les autres éléments ne devraient pas influencer négativement le ressenti des visiteurs car ils resteraient approximativement en face. Un problème que nous n'avons tout de même pas traité est comment obtenir l'apparence des hôtes sans leurs casques de réalité mixte (Gupta *et al.*, 2022; Zhao *et al.*, 2019). D'autres scénarios comme l'assistance à distance sont aussi intéressants à explorer. Par exemple, un technicien qui ramène sur site un expert à distance pour l'aider à réparer une machine. Avec la supposition d'un plus grand sentiment de présence que les applications actuelles, le visiteur expert devrait comprendre plus rapidement le problème du technicien. Dans ce cas, il est possible que l'incarnation en *billboard* ne soit plus suffisante et qu'il faille se pencher sur des incarnations volumétriques (Wang *et al.*, 2023a; Saito *et al.*, 2020, 2019). Si aujourd'hui la plupart des méthodes ne sont pas capables de générer ces incarnations en temps réel, on peut espérer que ce problème technique soit résolu dans le futur.

D'autres perspectives concernent le dispositif d'acquisition. Pour notre système idéal, nous avons argumenté que nous souhaitions baser notre système de télé-immersion sur une unique caméra omnidirectionnelle, notamment pour avoir un système nomade. Mais l'utilisation d'une unique caméra impose des limites indépassables sur la reconstruction 3D. Avec plusieurs caméras omnidirectionnelles, la qualité de la géométrie est améliorée et des occultations sont levées en couvrant des régions qui ne sont pas visibles avec une seule caméra. En terme topologique, cela implique des lieux qui ne sont pas des convexes étoilés pourraient être représentés plus fidèlement. Un futur système nomade pourrait alors relaxer cette contrainte et se baser sur plusieurs caméras omnidirectionnelles. Une autre possibilité pour conserver l'utilisation d'une unique caméra est de suivre une approche multi-modales, c'est-à-dire d'utiliser d'autres sources d'informations pour reconstituer la scène. Typiquement dans notre système, les hôtes portent des casques de réalité mixte pour voir l'incarnation des visiteurs. Or, ces dispositifs

sont généralement équipés de caméras pour assurer certaines fonctions comme la localisation (*inside-out tracking*). En supposant que les flux vidéos soient exploitables, ces caméras peuvent être utilisées conjointement avec la caméra omnidirectionnelle pour reconstruire les éléments de la scène. Ces caméras égocentriques sont particulièrement adaptées pour améliorer l'incarnation des hôtes, par exemple pour avoir des images de l'hôte sans son casque porté (Jourabloo *et al.*, 2022). Ces caméras étant aussi utilisées dans les casques modernes pour le suivi des mains, on peut alors imaginer estimer la pose de l'avatar d'un hôte conjointement avec la vidéo omnidirectionnelle et les vidéos embarquées. Enfin, toujours dans cette vision d'acquisition égocentrique, un futur système de télé-immersion peut reposer sur une caméra omnidirectionnelle embarquée par un hôte. Nous avons discuté que ce système n'était pas forcément idéal, car le point de vue est asservi à un hôte ou qu'il est difficile de retrouver des informations visuelles sur la personne qui porte la caméra. Mais le développement de nouvelles méthodes égocentriques lève ces contraintes. Ces méthodes permettent d'avoir un système de télé-immersion 3D 360° embarqué où le visiteur est directement immergé sur la zone d'intérêt du lieu, avec la pose de l'hôte (Hori *et al.*, 2022; Wang *et al.*, 2021b) et son apparence (Hu *et al.*, 2021) estimées de manière égocentrique.

Pour conclure sur les perspectives, nous pensons qu'il est très probable que n'importe quelle représentation 3D 360° dans l'avenir, construite à partir d'une unique caméra omnidirectionnelle fixe, restera toujours objectivement de moins bonne qualité qu'une reconstruction 3D basée sur plusieurs caméras. Les approches récentes pour créer des scènes à partir d'images, comme le rendu neuronal par NeRF (Mildenhall *et al.*, 2021) ou le *gaussian splatting* (Kerbl *et al.*, 2023) peuvent améliorer la qualité de notre rendu. Mais elles ne permettent pas résoudre le problème fondamental : comment avoir une représentation 3D crédible à partir d'une seule vue 360° ? Un gap doit être franchi pour savoir comment construire une représentation 3D 360° dans des cas comme quand l'ensemble du lieu n'est pas visible. Nous avons argumenté que les méthodes d'estimation de la géométrie actuelles à partir d'une seule image omnidirectionnelle ne sont pas pleinement satisfaisantes pour la réalité virtuelle. À ce jour, ces techniques altèrent la reproduction du relief de certains éléments du lieu ou engendrent des anomalies telles que des distorsions, affectant négativement la sensation de l'utilisateur d'être sur un lieu physique. Néanmoins, nous soutenons que le manque de précision dans l'estimation de la géométrie n'est pas fondamentalement préjudiciable pour la télé-immersion. Les limites de l'intelligence artificielle dans l'estimation mono-vue de la géométrie s'expliquent car ces approches minimisent en général des erreurs géométriques par rapport à une vérité terrain, en introduisant aussi parfois des contraintes pour garantir des propriétés comme une reconstruction lisse. Or, la reconstruction 3D 360° n'est pas destinée à conduire des simulations physiques, mais à immerger de manière crédible un visiteur sur un lieu

physique. La pure minimisation objective des erreurs géométriques n'est pas la garantie qu'une reconstruction 3D paraisse plus crédible aux yeux d'un visiteur. Déterminer les critères de la perception humaine à optimiser est crucial pour obtenir des reconstructions 3D réalistes dans lesquels un visiteur se sent vraiment transporté sur un lieu physique. Une direction qui paraît fructueuse est alors d'étudier les phénomènes psychologiques pour passer des reconstructions physiquement proches de la réalité à des reconstructions orientées utilisateurs, perceptuellement proches de la réalité, des reconstructions *pseudo-géométriques*. Des phénomènes connus de la perception humaine sont déjà exploités en imagerie de synthèse pour réduire objectivement la qualité graphique sans affecter la qualité perçue du rendu. Par exemple, la cécité d'inattention est utilisée pour diminuer la qualité des éléments qui n'ont pas l'attention d'un utilisateur (Cater *et al.*, 2002). Un tel phénomène est pertinent pour modéliser les objets ambiants d'une scène avec des représentations dégradées à qui on pourrait donner une géométrie grossièrement estimée. La texture est aussi une information suffisante à un utilisateur pour qu'il infère la géométrie d'un environnement (Kim *et al.*, 2004). La courbure d'un élément de l'image omnidirectionnelle pourrait alors être perçue, même s'il est projeté sur une géométrie dégradée comme une surface plate. L'idéal serait d'étudier les situations dans lesquelles la géométrie ou la topologie d'une scène peuvent être altérées sans atténuer le ressenti subjectif d'un utilisateur comme sa présence. Des expériences restent donc à mener pour trouver comment mitiger l'effet des imprécisions de la reconstruction 3D 360°, intrinsèques à l'utilisation d'une unique caméra omnidirectionnelle, sur la perception d'un utilisateur.

Aujourd'hui, nous pensons que le grand obstacle pour la généralisation de la télé-immersion 3D, et plus spécifiquement la télé-immersion 3D 360°, est la difficulté d'obtenir des informations 3D avec des appareils grand public. La webcam est un périphérique courant dans une configuration bureautique standard, tandis que les logiciels de messagerie instantanée sont largement utilisés dans notre quotidien. La visioconférence est alors une partie intégrante de la collaboration à distance, à tel point que l'enseignement à distance est désormais assimilé aux cours en visioconférence. Avec une acquisition 360° en 3D qui reste principalement réservée à un public d'experts, les systèmes de télé-immersion 3D 360° ne peuvent pas encore s'imposer dans la vie courante. Les applications grand public actuelles manipulant des images omnidirectionnelles ne requièrent pas d'informations géométriques. L'intérêt de la géométrie dans une image omnidirectionnelle n'est donc pas répandu dans l'opinion générale. Pourtant, il existe bien des caméras 360° avec capture 3D, mais ces appareils ne sont pas commercialisés car les intérêts applicatifs ne sont pas évidents. Continuer la recherche pour la télé-immersion 3D 360° pourrait alors montrer l'intérêt de ces appareils auprès d'acteurs industriels et les inciter à les produire en série, ce qui ouvrirait ces technologies à un public plus

large. La potentielle généralisation de la télé-immersion ouvre tout de même des questions éthiques qui doivent être abordées. Des risques de dérives existent dans son usage pour l'enseignement, comme à chaque fois qu'un problème social ou sociétal est adressé par des solutions techniques. Il sera absolument nécessaire dans le futur de clarifier les cas où il est acceptable de recourir à la télé-immersion plutôt que de faire déplacer les étudiants et les professeurs sur les lieux d'enseignements.

Annexe A

Classification des Systèmes de Télé-Immersion

TABLE A.1 – Classification des systèmes de télé-immersion rencontrés

Système	Extraction		Inclusion	
	Acquisition lieu hôte	Acquisition lieu visiteur	Incarnation hôte	Incarnation visiteur
(Alvarez *et al.*, 2022)	outside-in		avatar	
AVT (Rhee *et al.*, 2020)	inside-out	non-visuelle	vidéo	avatar
(Anjos *et al.*, 2019)	outside-in		avatar	
Beam [1]	inside-out	outside-in	vidéo	hybride
Beaming (Steed *et al.*, 2012)	inside-out	outside-in	vidéo	hybride
BeThere (Sodhi *et al.*, 2013)	inside-out	outside-in	avatar	avatar
blue-c (Gross *et al.*, 2003)		outside-in		avatar
Bublr [2]		outside-in		vidéo
C1x6 (Beck *et al.*, 2013)		outside-in		avatar
ChameleonMask (Misawa et Rekimoto, 2015)	inside-out	outside-in	vidéo	vidéo
Cisco TelePresence System 3000	outside-in		vidéo	
CollaVR (Nguyen *et al.*, 2017)		non-visuelle		vidéo

1. `https://telepresence.awabot.com`
2. `https://bublr.co`

TABLE A.1 – Classification des systèmes de télé-immersion rencontrés (suite)

Système	Extraction		Inclusion	
	Acquisition lieu hôte	Acquisition lieu visiteur	Incarnation hôte	Incarnation visiteur
(Dasari *et al.*, 2023)		outside-in		vidéo
(Deng *et al.*, 2023)		outside-in		avatar
EDGAR (Ching *et al.*, 2016)	inside-out	outside-in	vidéo	robot
FarfetchFusion (Lee *et al.*, 2023)	outside-in		avatar	
FreeWalk (Nishimura *et al.*, 1998)		outside-in		vidéo
(Fritsche *et al.*, 2015)	inside-out	non-visuelle	vidéo	robot
GazeChat (He *et al.*, 2021)		non-visuelle		vidéo
Geminoid (H et F) (Nishio *et al.*, 2007)	inside-out	non-visuelle	vidéo	robot
Giant-Miniature Collaboration (Piumsomboon *et al.*, 2019)	inside-out	non-visuelle	vidéo	avatar
Gutsy-Avatar (Tobita, 2017)	inside-out	outside-in	vidéo	vidéo
Holobot (Kim *et al.*, 2023)	inside-out	non-visuelle	vidéo	avatar
HoloKinect (Siemonsma et Bell, 2022)	outside-in		avatar	
Holoportation (Orts-Escolano *et al.*, 2016)	outside-in		avatar	
Horizon Workrooms [3]		non-visuelle		avatar
I3DVC (Kauff et Schreer, 2002)	outside-in		avatar	
(Irlitti *et al.*, 2023)	outside-in	non-visuelle	avatar	avatar
(Jones *et al.*, 2009)	inside-out	outside-in	vidéo	avatar
JackIn Neck (Matsuda *et al.*, 2018)	inside-out	outside-in	vidéo	vidéo
(Kobayashi *et al.*, 2021)	outside-in		vidéo	

3. `https://forwork.meta.com/horizon-workrooms`

TABLE A.1 – Classification des systèmes de télé-immersion rencontrés (suite)

Système	Extraction		Inclusion	
	Acquisition lieu hôte	Acquisition lieu visiteur	Incarnation hôte	Incarnation visiteur
Kumospace [4]		outside-in		vidéo
(Kurillo et Bajcsy, 2013)		outside-in		avatar
(Lee *et al.*, 2018)	inside-out	inside-out	vidéo	vidéo
(Leithinger *et al.*, 2014)	outside-in		robot	
LiveMask (Misawa *et al.*, 2012a)	inside-out	outside-in	vidéo	hybride
Ma petite chérie (Misawa *et al.*, 2012b)	inside-out	outside-in	vidéo	hybride
(Maimone et Fuchs, 2011)		outside-in		avatar
MeBot (Adalgeirsson et Breazeal, 2010)	inside-out	outside-in	vidéo	hybride
MH-2 (Han *et al.*, 2018)	inside-out	outside-in	vidéo	robot
MMSpace (Otsuka, 2016)	outside-in		hybride	
Mobileportation (Young *et al.*, 2020)	inside-out	inside-out	vidéo	vidéo
(Nagendran *et al.*, 2015)	non-visuelle		robot	
(Ogi et Fueki, 2017)	inside-out	non-visuelle	vidéo	hybride
OmniGaze (Shiro *et al.*, 2018)	inside-out	outside-in	vidéo	vidéo
(Onishi *et al.*, 2016)	outside-in		robot	
OpenIMPRESS (Kolkmeier *et al.*, 2018)	outside-in	non-visuelle	avatar	avatar
(Paulos et Canny, 2001)	inside-out	outside-in	vidéo	hybride
(Pece *et al.*, 2013)		outside-in		vidéo
(Plüss *et al.*, 2016)	outside-in		avatar	
Polly (Kimber *et al.*, 2015)	inside-out	outside-in	vidéo	hybride
PopObject (Kushida et Nakanishi, 2018)	outside-in		robot	
RemoteCoDe (Sakashita *et al.*, 2022)	inside-out		hybride	
Room2Room (Pejsa *et al.*, 2016)	outside-in		avatar	

4. `https://kumospace.com`

TABLE A.1 – Classification des systèmes de télé-immersion rencontrés (suite)

Système	Extraction		Inclusion	
	Acquisition lieu hôte	Acquisition lieu visiteur	Incarnation hôte	Incarnation visiteur
Shader Lamps Avatar (Schubert *et al.*, 2012)	inside-out	outside-in	vidéo	hybride
(Sirkin *et al.*, 2011)	inside-out	outside-in	vidéo	hybride
Starline (Lawrence *et al.*, 2021)	outside-in		avatar	
(Tejwani *et al.*, 2023)	inside-out	outside-in	vidéo	robot
Teledrone (Shakeri et Neustaedter, 2019)	inside-out	outside-in	vidéo	hybride
TeleHuman2 (Gotsch *et al.*, 2018)	inside-out	outside-in	vidéo	avatar
Telenoid (Ogawa *et al.*, 2011)	inside-out	outside-in	vidéo	robot
TELEPORT (Gibbs *et al.*, 1999)	outside-in		avatar	
TELESAR (I à VI) (Tachi, 2019)	inside-out	non-visuelle	vidéo	robot
(Teo *et al.*, 2019)	inside-out	inside-out	vidéo	vidéo
TEROOS (Kashiwabara *et al.*, 2012)	inside-out	non-visuelle	vidéo	hybride
ThirdEye (Otsuki *et al.*, 2017)	inside-out	outside-in	vidéo	vidéo
(Tobita *et al.*, 2011)	inside-out	outside-in	vidéo	hybride
(Tokuda *et al.*, 2013)	inside-out	non-visuelle	vidéo	avatar
Tourgether360 (Kumar *et al.*, 2022)		non-visuelle		avatar
(Towles *et al.*, 2003)	outside-in		avatar	
Viewpoint-Controllable Telepresence (Luo *et al.*, 2023)	inside-out	non-visuelle	vidéo	avatar
VirtualCube (Zhang *et al.*, 2022b)	outside-in		avatar	
Visioconférence		outside-in		vidéo

TABLE A.1 – Classification des systèmes de télé-immersion rencontrés (suite)

Système	Extraction		Inclusion	
	Acquisition lieu hôte	Acquisition lieu visiteur	Incarnation hôte	Incarnation visiteur
Vivid (Hopkins et Benford, 1996)		outside-in		vidéo
VROOM (Jones *et al.*, 2020a)	inside-out	non-visuelle	vidéo	avatar
WACL (Sakata *et al.*, 2003)	inside-out	non-visuelle	vidéo	hybride
(Wang *et al.*, 2023b)	outside-in		avatar	
withyou (Roberts *et al.*, 2015)		outside-in		avatar
(Young *et al.*, 2019)	inside-out	outside-in	vidéo	vidéo
(Yu *et al.*, 2021)	outside-in	non-visuelle	avatar	avatar
(Zioulis *et al.*, 2016)		outside-in		avatar

L'acquisition d'un lieu hôte correspond au dispositif d'acquisition utilisé sur le lieu d'un hôte. L'acquisition d'un lieu visiteur correspond au dispositif d'acquisition utilisé sur le lieu d'un visiteur. Un dispositif outside-in capture généralement un environnement tandis qu'un dispositif inside-out capture un objet. Les cellules des systèmes où le dispositif d'acquisition n'est pas visuel sont notées non-visuelles. L'incarnation d'un hôte correspond à la manière dont un utilisateur perçoit un hôte. L'incarnation d'un visiteur correspond à la manière dont un utilisateur perçoit un visiteur. Les systèmes sans incarnations n'ont pas été inclus. Les cellules en rapport avec l'hôte ou le visiteur sont laissées vides si le système n'a pas d'utilisateur jouant ce rôle. Les systèmes symétriques se distinguent alors par le fait qu'ils possèdent des attributs exclusivement liés à l'hôte ou exclusivement liés au visiteur.

Annexe B

Protocole Expérimentale, Formulaire de Consentement et Questionnaires

Protocole d’expérimentation

LSI - CEA-List - Paris Saclay

G-SCOP - Grenoble INP - Université de Grenoble

Résumé du projet

Ces dernières années, les périodes vécues de confinement ont transformé de nombreuses pratiques. En particulier, pour maintenir la formation des étudiants, cette période a vu la mise en place de l'enseignement à distance se généraliser. Dans sa configuration la plus simple, les étudiants se retrouvent sur une plateforme de visioconférence type zoom [1] où le professeur fait cours avec le flux vidéo de la webcam. Cette nouvelle manière d'enseigner a de nombreux inconvénients comparé aux cours donnés en présentiel. Par exemple, les professeurs peuvent être moins conscients de l'état de concentration des étudiants. Les étudiants, se sentant moins impliqués avec juste un retour de la webcam du professeur, peuvent perdre en concentration et moins retenir les informations. Aussi, les interactions individuelles entre les étudiants sont limitées, ce qui complique les échanges entre eux. Enfin, la tenue de travaux pratique où il y a besoin de manipuler des éléments n'est pas concevable avec une interface de visioconférence limitée. Néanmoins, dans de nombreux cas, il peut être intéressant de développer des outils plus élaborés pour tenir des cours à distance pour palier à l'impossibilité des étudiants de se déplacer ou à l'incapacité d'accueil de l'établissement. Une piste pour régler les problèmes des cours en visioconférence serait d'utiliser la réalité virtuelle. Ces technologies immersives permettraient idéalement d'immerger professeurs et étudiants dans une salle de classe et d'avoir un cours comme s'il se déroulait en présentiel. Combinés avec des jumeaux numériques, les professeurs pourraient aussi développer des sessions de travaux pratiques où un étudiant pourrait manipuler une copie virtuelle d'une machine comme s'il la manipulait en réel. C'est dans ce contexte qu'a été développé le projet JENII (Jumeaux d'Enseignement Numériques, Immersifs et Interactifs) qui vise à créer des formations à distance immersive et collaborative, dans lequel le CEA est impliqué avec les Arts et Métiers, le CESI et le CNAM.

Nous explorons la piste utilisant une caméra 360° au lieu d'une simple webcam pour filmer le professeur donnant cours dans la salle de classe. Cette méthode permet d'immerger en réalité virtuelle les étudiants à distance dans la salle de classe comme s'ils étaient en présentiel. Mais cette solution conserve un problème, la salle de classe n'étant visible que du point de vue de la caméra 360°, les étudiants ne peuvent pas se déplacer et donc ne peuvent pas se voir les uns les autres. Une solution technique que nous avons proposée est d'utiliser des méthodes d'apprentissage profond afin de réaliser une reconstruction 3D basé sur le point de vue de la caméra 360°. L'objectif de cette expérimentation consiste à montrer l'apport de notre proposition de reconstruction 3D comparé à une vidéo 360° traditionnelle dans le cadre de l'enseignement à distance.

Titre du projet

Évaluation de l'apport de la reconstruction 3D pour la télé-immersion basé caméra 360°

Domaine scientifique

Réalité virtuelle, Vidéo 360°, Reconstruction 3D

Responsable scientifique du projet

DLUZNIEWSKI Clément
clement.dluzniewski@cea.fr
Doctorant CEA-List - G-SCOP

Participants au projet

Étudiants en réalité virtuelle à Grenoble-INP et membres du LSI au CEA.

Lieu(x) de recherche où l'étude va être conduite

Salle de réalité virtuelle au G-SCOP et bureau du LSI. L'expérimentation se déroulant dans un environnement virtuelle, le lieu où l'étude est conduite ne devrait pas avoir d'incidence.

Objectif principal

Évaluer l'apport des méthodes de reconstruction 3D par rapport à la vidéo 360° classique pour la collaboration entre des personnes sur site et à distance.

1 – Description du projet

Contexte et intérêt scientifique

Le projet s'inscrit dans le cadre du projet JENII (Jumeaux d'Enseignement Numériques, Immersifs et Interactifs) qui développe l'utilisation d'environnements immersifs et collaboratifs pour la formation à distance. Une des propositions avancées est d'utiliser une caméra 360° pour capturer une formation et de la retransmettre en temps réel à des utilisateurs distants en réalité virtuelle. Cette proposition se justifie par le fait que cette méthode est plus immersive que visualiser le cours à travers un écran filmé par une caméra traditionnelle. L'étudiant se sentant comme s'il était présent dans la salle de formation, celui-ci devrait mieux retenir les informations qu'avec des approches non-immersives. Cependant, les technologies de réalité virtuelle utilisées actuellement ne permettent pas à un utilisateur qui visualise une vidéo 360° de se déplacer librement dans l'environnement virtuel. Pour résoudre ce problème, nous avons proposé des méthodes basées sur de l'apprentissage profond afin de réaliser une reconstruction 3D basé sur le point de vue de la caméra 360°. Cette étude vise à évaluer l'apport de notre approche de reconstruction 3D par rapport à la vidéo 360° classique sur des critères subjectifs.

Objectifs

L'objectif est de montrer que l'ajout de la 3D, avec une carte de profondeur pour le fond et une représentation billboard pour les personnes, permet un plus grand sentiment de présence qu'avec une vidéo 360° classique sans 3D. La présence spatiale et la présence sociale en particulier seront évaluées. Ces différents aspects de la télé-présence seront mesurés avec le questionnaire *Temple Presence Inventory* (TPI) [2]. L'utilisabilité et la charge de travail du système seront aussi évaluées avec les questionnaires *System Usability Scale* (SUS) [3] et le NASA-TLX [4].

Hypothèses de recherche

Des expériences réalisées précédemment semblent indiquer que la possibilité de contrôler le point de vue et la présence d'une parallaxe de mouvement améliore le sentiment de présence de l'utilisateur. L'hypothèse de recherche est alors que l'ajout de la 3D à la vidéo 360° permet d'augmenter la présence spatiale et la présence sociale. On suppose alors que la présence d'artefacts de reconstruction 3D, visibles lorsque le participant s'éloigne de la position de la caméra, ne dégradera pas le sentiment de présence autant que l'interactivité ajoutée par la 3D contribue à l'augmenter. On suppose aussi que l'utilisabilité et la charge de travail seront les mêmes avec et sans 3D.

2 – Matériel et méthodes

A – Participants

Recrutement

Mode de recrutement : Les participants seront principalement recrutés dans le cadre de session de travaux pratiques sur la réalité virtuelle. Des membres du LSI seront aussi sollicités individuellement.

Critères de sélection : Aucun

Critère de non-inclusion : Le participant devant suivre une conversation en anglais dans une vidéo et répondre aux questionnaires en anglais, une compréhension de l'anglais est nécessaire.

Indemnisation

Aucune

B – Méthode

Description du protocole

Le scenario de l'expérience consiste pour le participant à visualiser une vidéo 360°, sans ou avec 3D. La vidéo 360° est une vidéo de personnes dans une salle donnant oralement des explications sur un système. Deux modalités seront testées pour la représentation de la vidéo 360° : 360° et 3D. La modalité 360° consiste à visualiser la vidéo 360° en réalité virtuelle avec un outil de visualisation de vidéo 360° classique, c'est-à-dire avec uniquement la possibilité de changer l'orientation de la tête, mais sans possibilité de se déplacer. La modalité 3D consiste à visualiser la vidéo 360° en réalité virtuelle avec nos approches de reconstruction 3D. La représentation 3DV de la vidéo en 3D consiste à ajouter du relief sur le fond statique de la vidéo 360° (grâce à une carte de profondeur) et à utiliser des billboards pour représenter les personnes. Dans ce cas, le participant peut se déplacer librement dans l'environnement virtuel.
À l'arrivée du participant, celui-ci est accueilli avec le formulaire d'information et de consentement à la participation. Il remplira aussi un formulaire sociodémographique (questionnaire préliminaire) afin de mieux le connaitre. L'expérience est conçue de manière intra-sujets : chaque participant testera les deux modalités (360° ou 3D). L'ordre d'exposition aux modalités est choisi aléatoirement. Dans les deux modalités, le participant devra rester debout et sera immergé dans l'environnement virtuel avec un casque de réalité virtuelle. On

initialise l'environnement avec la première image de la vidéo et on laisse le participant explorer la scène avant de démarrer la vidéo. Avant la vidéo avec la modalité 360°, le participant sera incité à regarder autour de lui (impossibilité de se déplacer dans l'environnement virtuel). Avec la modalité 3D, le participant sera incité à se déplacer dans l'environnement virtuel. La salle d'expérimentation devra alors être suffisamment grande pour que le participant puisse se déplacer librement. La vidéo est lancée quand le participant est prêt. Après avoir été immergé dans la vidéo, le participant remplit les questionnaires d'évaluation. Les conditions 360° et 3D seront appelées respectivement A et B auprès des participants.

Matériels utilisés

Casque de réalité virtuelle relié à un ordinateur. Si le casque est connecté à un ordinateur en filaire, prévoir un câble suffisamment long pour que les participants puissent se déplacer sans contraintes.

Calendrier des expérimentations

Période de tenue des travaux pratiques sur la réalité virtuelle en génie industriel à Grenoble-INP (octobre 2023 à décembre 2023).

Analyse des données

Les données à analyser seront extraites des réponses des questionnaires posés après avoir visualisé la scène avec la modalité testée. L'analyse consistera à montrer que la présence est plus grande dans la condition 3D que dans la condition 360°, et que l'utilisabilité et la charge de travail seront équivalentes dans les deux conditions. L'analyse veillera aussi à ce que l'ordre d'exposition n'ait pas d'effet sur la présence, l'utilisabilité et la charge de travail.

C – Bénéfices et risques prévisibles et connus pour la santé physique et mentale (estime de soi, etc.) et la vie sociale (réputation)

Aucun

D – Vigilance et arrêt prématuré de l'étude

Critères d'arrêt de l'étude pour un participant

Un participant peut décider d'arrêter sa participation à tout moment quelles qu'en soient ses raisons.

3 – Traitement des données et respect de la vie privée du participant

A – Confidentialité

Procédé d'anonymisation

Les réponses aux questionnaires enregistrées seront anonymes. Il n'existe pas de correspondance entre les données enregistrées et l'identité d'un participant.

Personnes ayant accès aux données

DLUZNIEWSKI Clément, NOËL Frédéric, LE GARREC Jérémie, ANDRIOT Claude

B – Archivage

Type de données archivées (préciser si données identifiantes, directement ou par recoupement)

Questionnaire sociodémographique : age, sexe, niveau d'étude, maitrise de l'anglais, vision, familiarité avec la réalité virtuelle, familiarité avec la vidéo 360°
Questionnaire sur une modalité : modalité testée (A / B), ordre de passage (A-B / B-A), résultats du TPI, résultats du SUS, résultats du NASA-TLX
Les données saisies par un participants et enregistrées ne permettent pas de retrouver son identité.

Durée d'archivage

10 ans

Lieu d'archivage

G-SCOP - Bureau de NOËL Frédéric

Personne responsable

NOËL Frédéric - Laboratoire G-SCOP

Possibilité de destruction à la demande du participant

L'anonymat rend impossible la rectification ou la suppression des informations concernant un participant après la fin de leur participation.

Références

[1] https://zoom.us/
[2] M. Lombard, T. Bolmarcich, et L. Weinstein, *Measuring Presence: The Temple Presence Inventory*, 2009
[3] J. Brooke, *SUS - a quick and dirty usability scale*, 1996
[4] S. G. Hart et S. E. Lowell, *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*, 1988

Information and participation consent form

Given to those invited to take part in this research experiment

Principal investigator

Clément Dluzniewski - Doctoral student

Research supervisor

Frédéric Noël - Research supervisor

Place of experimentation

G-SCOP, Conception Collaborative team, Building R, Room R006
46 avenue Félix Viallet 38013 Grenoble Cedex 1 - France

Objective

Evaluate the possibilities of creating a virtual reality scene from a 360° video, and their impact on presence, usability and workload.

Methodology

We want to compare different existing modalities for immersing a user in a virtual reality scene created from a 360° video.
For this purpose, you will be invited to watch a 360° video in virtual reality under two different conditions. In one condition you'll be able to move freely around the scene, while in the other you'll be in a fixed position. The order of the conditions will be randomly assigned. Before each video you will be immersed in an image corresponding to the condition being tested to familiarize yourself with the controls. After each condition, you'll fill in a questionnaire to evaluate the system in terms of presence, usability and workload.

You can stop the experiment at any time if you request it.

I have read and understood the instructions and I am willing to participate in this experiment.
I give my consent to allow the laboratory to use the data collected for research purposes only for the duration of the principal investigator studies.

Done at …….. on ……..

Signature :

Preliminary form

1. What is your age ?

2. What is your sex ?

   Mark only one oval.

   - ◯ Male
   - ◯ Female

3. What is your level of education ?

   Mark only one oval.

   - ◯ Baccalauréat or lower
   - ◯ DUT, BTS, DEUG (Bac+2)
   - ◯ Licence (Bac+3)
   - ◯ Master (Bac+4, Bac+5)
   - ◯ Doctorat (Bac+8)

4. Is english your native language ?

   Mark only one oval.

   - ◯ Yes
   - ◯ No

5. Do you have eyesight problems ?

Mark only one oval.

- Yes
- No

6. If concerned, are you wearing vision correction equipment for this experience ?

Mark only one oval.

- Not concerned
- Yes
- No

7. How familiar are you with virtual reality devices ?

Mark only one oval.

	1	2	3	4	5	6	7	
Not familiar								Very familiar

8. How familiar are you with 360° videos ?

Mark only one oval.

	1	2	3	4	5	6	7	
Not familiar								Very familiar

Post-experience evaluation questionnaire

This questionnaire assesses the sense of presence, the usability and the workload of the system.

* Indicates required question

1. In what order do you test the conditions ? *

   Mark only one oval.

   - ◯ A-B
   - ◯ B-A

2. What conditions have you experienced ? *

   Mark only one oval.

   - ◯ A
   - ◯ B

Temple Presence Inventory

Spatial presence

3. How much did it seem as if the objects and people you saw/heard had come to the place you were ? *

   Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

4. How much did it seem as if you could reach out and touch the objects or people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

5. How often when an object seemed to be headed toward you did you want to move to get out of its way ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Never	◯	◯	◯	◯	◯	◯	◯	Always

6. To what extent did you experience a sense of being there inside the environment you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

7. To what extent did it seem that sounds came from specific different locations ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

8. How often did you want to or try to touch something you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Never	◯	◯	◯	◯	◯	◯	◯	Always

Social presence - Actor within medium (parasocial interaction)

9. How often did you have the sensation that people you saw/heard could also see/hear you ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Never	◯	◯	◯	◯	◯	◯	◯	Always

10. To what extent did you feel you could interact with the person or people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
None	◯	◯	◯	◯	◯	◯	◯	Very much

11. How much did it seem as if you and the people you saw/heard both left the places where you were and went to a new place ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

12. How much did it seem as if you and the people you saw/heard were together in the same place ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

13. How often did it feel as if someone you saw/heard in the environment was talking directly to you ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Never	◯	◯	◯	◯	◯	◯	◯	Always

14. How often did you want to or did you make eye-contact with someone you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Never	◯	◯	◯	◯	◯	◯	◯	Always

15. Seeing and hearing a person through a medium constitutes an interaction with him or her. How much control over the interaction with the person or people you saw/heard did you feel you had ? *

Mark only one oval.

	1	2	3	4	5	6	7	
None	◯	◯	◯	◯	◯	◯	◯	Very much

Social presence - Passive interpersonal

16. During the media experience how well were you able to observe the facial expressions of the people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not well								Very well

17. During the media experience how well were you able to observe the changes in tone of voice of the people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not well								Very well

18. During the media experience how well were you able to observe the style of dress of the people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not well								Very well

19. During the media experience how well were you able to observe the body language of the people you saw/heard ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not well								Very well

Engagement (mental immersion)

20. To what extent did you feel mentally immersed in the experience ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all								Very much

21. How involving was the experience ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all								Very much

22. How completely were your senses engaged ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all								Very much

23. To what extent did you experience a sensation of reality ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all								Very much

24. How relaxing or exciting was the experience ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very relaxing	◯	◯	◯	◯	◯	◯	◯	Very exciting

25. How engaging was the story ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all	◯	◯	◯	◯	◯	◯	◯	Very much

Social richness

Please choose the number that best describes your evaluation of the media experience

26. *

Mark only one oval.

	1	2	3	4	5	6	7	
Remote	◯	◯	◯	◯	◯	◯	◯	Immediate

27. *

Mark only one oval.

	1	2	3	4	5	6	7	
Unemotional	◯	◯	◯	◯	◯	◯	◯	Emotional

28. *

Mark only one oval.

	1	2	3	4	5	6	7	
Unresponsive	◯	◯	◯	◯	◯	◯	◯	Responsive

29. *

Mark only one oval.

	1	2	3	4	5	6	7	
Dead	◯	◯	◯	◯	◯	◯	◯	Lively

30. *

Mark only one oval.

	1	2	3	4	5	6	7	
Impersonal	◯	◯	◯	◯	◯	◯	◯	Personal

31. *

Mark only one oval.

	1	2	3	4	5	6	7	
Insensitive	◯	◯	◯	◯	◯	◯	◯	Sensitive

32. *

Mark only one oval.

	1	2	3	4	5	6	7	
Unsociable	◯	◯	◯	◯	◯	◯	◯	Sociable

System usability scale

33. I think that I would like to use this system frequently. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

34. I found the system unnecessarily complex. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

35. I thought the system was easy to use. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

36. I think that I would need the support of a technical person to be able to use this system. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

37. I found the various functions in this system were well integrated. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

38. I thought there was too much inconsistency in this system. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

39. I would imagine that most people would learn to use this system very quickly. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

40. I found the system very cumbersome to use. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

41. I felt very confident using the system. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

42. I needed to learn a lot of things before I could get going with this system. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

NASA Task Load Index

43. How mentally demanding was the task ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very low	◯	◯	◯	◯	◯	◯	◯	Very high

44. How physically demanding was the task ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very low	◯	◯	◯	◯	◯	◯	◯	Very high

45. How hurried or rushed was the pace of the task ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very low	◯	◯	◯	◯	◯	◯	◯	Very high

46. How successful were you in accomplishing what you were asked to do ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Perfect	◯	◯	◯	◯	◯	◯	◯	Failure

47. How hard did you have to work to accomplish your level of performance ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very low	◯	◯	◯	◯	◯	◯	◯	Very high

48. How insecure, discouraged, irritated, stressed, and annoyed were you ? *

Mark only one oval.

	1	2	3	4	5	6	7	
Very low	◯	◯	◯	◯	◯	◯	◯	Very high

Bibliographie

Sigurdur Orn Adalgeirsson et Cynthia Breazeal : MeBot : A robotic platform for socially embodied telepresence. *In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 15–22, 2010.

Matt Adcock, Stuart Anderson et Bruce Thomas : RemoteFusion : real time depth camera fusion for remote collaboration on physical tasks. *In Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '13, pages 235–242. Association for Computing Machinery, 2013.

Marina Alvarez, Alexander Toet et Sylvie Dijkstra-Soudarissanane : Virtual Visits : UX Evaluation of a Photorealistic AR-based Video Communication Tool. *In Proceedings of the 1st Workshop on Interactive eXtended Reality*, IXR '22, pages 69–75. Association for Computing Machinery, 2022. ISBN 978-1-4503-9501-4.

Terry Anderson et Pablo Rivera Vargas : A Critical look at Educational Technology from a Distance Education Perspective. *Articles publicats en revistes (Didàctica i Organització Educativa)*, 2020. ISSN 2013-9144.

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent et Josh Weaver : Google Street View : Capturing the World at Street Level. *IEEE Computer*, 43:32–38, 2010.

Rafael Anjos, Mauricio Sousa, Daniel Mendes, Daniel Medeiros, Mark Billinghurst, Craig Anslow et Joaquim Jorge : Adventures in Hologram Space : Exploring the Design Space of Eye-to-eye Volumetric Telepresence. *In Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, pages 1–5, 2019. ISBN 978-1-4503-7001-1.

Iqra Arshad, Paulo De Mello, Martin Ender, Jason D. McEwen et Elisa R. Ferré : Reducing Cybersickness in 360-Degree Virtual Reality. *Multisensory Research*, 35 (2):203–219, 2021. ISSN 2213-4808, 2213-4794.

Kohei Aso, Dong-Hyun Hwang et Hideki Koike : Portable 3D Human Pose Estimation for Human-Human Interaction using a Chest-mounted Fisheye Camera. *In Augmented Humans Conference 2021*, AHs'21, pages 116–120. Association for Computing Machinery, 2021. ISBN 978-1-4503-8428-5.

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt et James Tompkin : MatryODShka : Real-time 6DoF Video View Synthesis using Multi-Sphere Images. *In Computer Vision - ECCV 2020*, Lecture Notes in Computer Science, pages 441–459. Springer International Publishing, 2020. ISBN 978-3-030-58452-8.

Lucio Azzari, Federica Battisti et Atanas Gotchev : Comparative analysis of occlusion-filling techniques in depth image-based rendering for 3D videos. *In Proceedings of the 3rd workshop on Mobile video delivery*, MoViD '10, pages 57–62. Association for Computing Machinery, 2010. ISBN 978-1-4503-0165-7.

Jeremy N. Bailenson : Nonverbal Overload : A Theoretical Argument for the Causes of Zoom Fatigue. *Technology, Mind, and Behavior*, 2(1), 2021. ISSN 2689-0208.

Liam Bannon : Perspectives on CSCW : From HCI and CMC to CSCW. *In EWHCI'92 : Proc. Int. Conf. on HCI, August 1992, St. Petersburg, Russia*, pages 148–158, 1992.

Woodrow Barfield, Claudia Hendrix et Karl Bystrom : Visualizing the structure of virtual objects using head tracked stereoscopic displays. *In Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*, pages 114–120, 1997.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan et Peter Hedman : Mip-NeRF 360 : Unbounded Anti-Aliased Neural Radiance Fields. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469. IEEE Computer Society, 2022. ISBN 978-1-66546-946-3.

Robert Bauernschmitt, Martin Buss, Barbara Deml, Klaus Diepold, Berthold Färber, Georg Färber, Ulrich A. Hagn, Gerd Hirzinger, Sandra Hirche, Alois Knoll, Hermann Müller, Tobias Ortmaier, Angelika Peer, Michael Popp, Carsten Preusche, Gunther Reinhart, Zhuanghua Shi, Eckehard Steinbach, Heinz Ulbrich, Ulrich Walter et Michael F. Zäh : High-fidelity telepresence and teleaction. *In 2010 IEEE International Conference on Robotics and Automation*, pages 1092–1093, 2010.

André Bazin : *Qu'est-ce que le cinéma ?* Cerf, 1976. ISBN 978-2-204-02419-8.

Stephan Beck, André Kunert, Alexander Kulik et Bernd Froehlich : Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013. ISSN 1941-0506.

Brooke Belisle : Nature at a glance : Immersive maps from panoramic to digital. *Early Popular Visual Culture*, 13:313–335, 2015.

Ryad Benosman et Sing Bing Kang : *A Brief Historical Perspective on Panorama*, pages 5–20. Monographs in Computer Science. Springer, 2001. ISBN 978-1-4757-3482-9.

Tobias Bertel, N. D. F. Campbell et Christian Richardt : MegaParallax : Casual 360° Panoramas with M]otion Parallax. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

Tobias Bertel, Mingze Yuan, Reuben Lindroos et Christian Richardt : OmniPhotos : casual 360° VR photography. *ACM Transactions on Graphics*, 39(6):266 :1–266 :12, 2020. ISSN 0730-0301.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka et Matthias Müller : ZoeDepth : Zero-shot Transfer by combining Relative and Metric Depth. *arXiv*, 2023.

Yuri Antonio Gonçalves Vilas Boas : Overview of virtual reality technologies. *In Interactive Multimedia Conference*, 2013.

Leanne S. Bohannon, Andrew M. Herbert, Jeff B. Pelz et Esa M. Rantanen : Eye contact and video-mediated communication : A review. *Displays*, 34(2):177–185, 2013. ISSN 0141-9382.

Mehdi Boukhris, Alexis Paljic et Dominique Lafon-Pham : 360° versus 3D Environments in VR Headsets for an Exploration Task. *ICAT-EGVE 2017 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 2017. ISSN 1727-530X.

Jérôme Bourdon : Rethinking telepresence : post- and pre-covid-19. *Media, Culture & Society*, 45, 2023.

Douglas A. Bowman, Ernst Kruijff, Joseph J. LaViola et Ivan Poupyrev : An Introduction to 3-D User Interface Design. *Presence : Teleoperators and Virtual Environments*, 10(1):96–108, 2001.

Douglas A. Bowman et Ryan Patrick McMahan : Virtual Reality : How Much Immersion Is Enough ? *Computer*, 40(7):36–43, 2007. ISSN 1558-0814.

Aras Bozkurt : From distance education to open and distance learning : A holistic evaluation of history, definitions, and theories. *In Handbook of Research on Learning in the Age of Transhumanism*, pages 252–273. IGI Global, 2019.

Eleonora Brivio, Silvia Serino, Erica Negro Cousa, Andrea Zini, Giuseppe Riva et Gianluca De Leo : Virtual reality and 360° panorama technology : a media comparison to study changes in sense of presence, anxiety, and positive emotions. *Virtual Reality*, 25(2):303–311, 2021. ISSN 1434-9957.

john Brooke : SUS : A "Quick and Dirty" Usability Scale. *Usability Evaluation In Industry*, 189(194):4–7, 1996.

Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen et Paul Debevec : Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics*, 39(4):86 :86 :1–86 :86 :15, 2020. ISSN 0730-0301.

Angelika Bullinger-Hoffmann, Michael Koch, Kathrin Möslein et Alexander Richter : Computer-Supported Cooperative Work - Revisited. *i-com*, 20(33):215–228, 2021. ISSN 2196-6826.

Vannevar Bush : As we may think. *The atlantic monthly*, 176(1):101–108, 1945.

Davide Calandra, F. Gabriele Pratticò, Alberto Cannavò, Claudio Casetti et Fabrizio Lamberti : Digital twin- and extended reality-based telepresence for collaborative robot programming in the 6G perspective. *Digital Communications and Networks*, 2022. ISSN 2352-8648.

Abraham Campbell, Thomas Holz, Jonny Cosgrove, Mike Harlick et Tadhg O'Sullivan : Uses of Virtual Reality for Communication in Financial Services : A Case Study on Comparing Different Telepresence Interfaces : Virtual Reality Compared to Video Conferencing. *In Lecture Notes in Networks and Systems*, pages 463–481, 2020. ISBN 9789811336232.

Chao Cao, Marius Preda, Vladyslav Zakharchenko, Euee S. Jang et Titus Zaharia : Compression of Sparse and Dense Dynamic Point Clouds-Methods and Standards. *Proceedings of the IEEE*, 109(9):1537–1558, 2021. ISSN 1558-2256.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh et Jason Saragih : Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics*, 41(4):163 :1–163 :19, 2022. ISSN 0730-0301.

Kirsten Cater, Alan Chalmers et Patrick Ledda : Selective quality rendering by exploiting human inattentional blindness : looking but not seeing. *In Proceedings of the ACM symposium on Virtual reality software and technology*, VRST '02, pages 17–24. Association for Computing Machinery, 2002. ISBN 978-1-58113-530-5.

Vivian Chan, Nathaniel D Larson, David A Moody, David G Moyer et Neeral L Shah : Impact of 360° vs 2D Videos on Engagement in Anatomy Education. *Cureus*, 13(4), 2021. ISSN 2168-8184.

Yuan Chang et Guo-Ping Wang : A review on image-based rendering. *Virtual Reality & Intelligent Hardware*, 1(1):39–54, 2018. ISSN 2096-5796.

Shenchang Eric Chen : QuickTime VR : an image-based approach to virtual environment navigation. *In Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '95, pages 29–38. Association for Computing Machinery, 1995. ISBN 978-0-89791-701-8.

Zhenzhong Chen, Yiming Li et Yingxue Zhang : Recent advances in omnidirectional video coding for virtual reality : Projection and evaluation. *Signal Processing*, 146:66–78, 2018. ISSN 0165-1684.

Ho Kei Cheng et Alexander G. Schwing : XMem : Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. *In Computer Vision - ECCV 2022*, pages 640–658. Springer Nature Switzerland, 2022.

Pang Wee Ching, Wong Choon Yue et Gerald Seet Gim Lee : Design and Development of EDGAR - A Telepresence Humanoid for Robot-mediated Communication and Social Applications. *In 2016 IEEE International Conference on Control and Robotics Engineering (ICCRE)*, pages 1–4, 2016.

SungIk Cho, Seung-wook Kim, JongMin Lee, JeongHyeon Ahn et JungHyun Han : Effects of volumetric capture avatars on social presence in immersive virtual environments. *In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 26–34, 2020.

Miriam Clemente, Alejandro Rodríguez, Beatriz Rey et Mariano Alcañiz : Assessment of the influence of navigation control and screen size on the sense of presence in virtual reality using eeg. *Expert Systems with Applications*, 41(4, Part 2):1584–1592, 2014. ISSN 0957-4174.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk et Steve Sullivan : High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69 :1–69 :13, 2015. ISSN 0730-0301.

Xavier Corbillon : *Enable the next generation of interactive video streaming.* Thèse de doctorat, Ecole nationale supérieure Mines-Télécom Atlantique, 2019.

Diana-Margarita Córdova-Esparza, Juan R. Terven, Hugo Jiménez-Hernández, Ana Herrera-Navarro, Alberto Vázquez-Cervantes et Juan-M. García-Huerta : Low-bandwidth 3D visual telepresence system. *Multimedia Tools and Applications*, 78(15):21273–21290, 2019. ISSN 1573-7721.

Thiago L. T. da Silveira et Claudio R. Jung : Dense 3D Scene Reconstruction from Multiple Spherical Images for 3-DoF+ VR Applications. *In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 9–18, 2019.

Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena et Cláudio R. Jung : 3D Scene Geometry Estimation from 360° Imagery : A Survey. *ACM Computing Surveys*, 2022. ISSN 0360-0300.

Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi et Daniele Pucci : Teleoperation of Humanoid Robots : A Survey. *IEEE Transactions on Robotics*, 39(3):1706–1727, 2023. ISSN 1941-0468.

Mallesham Dasari, Edward Lu, Michael W. Farb, Nuno Pereira, Ivan Liang et Anthony Rowe : Scaling VR Video Conferencing. *In 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 648–657, 2023.

Henrique Galvan Debarba, Mario Montagud, Sylvain Chagué, Javier Garcia-Lajara Herrero, Ignacio Lacosta, Sergi Fernandez Langa et Caecilia Charbonnier : Content format and quality of experience in virtual reality. *Multimedia Tools and Applications*, 2022. ISSN 1573-7721.

T. DeFanti, D. Sandin, G. Dawe, M. Brown, M. Rawlings, G. Lindahl, A. Johnson et J. Leigh : Personal Tele-immersion devices. *In The Seventh International Symposium on High Performance Distributed Computing*, pages 198–205, 1998.

Haoke Deng, Qimeng Zhang, Hongyu Jin et Chang-Hun Kim : Real-Time Interaction for 3D Pixel Human in Virtual Environment. *Applied Sciences*, 13(22):966, 2023. ISSN 2076-3417.

Daniel DeTone, Tomasz Malisiewicz et Andrew Rabinovich : SuperPoint : Self-Supervised Interest Point Detection and Description. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 337–33712. IEEE, 2018.

Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab et Federico Tombari : Peeking behind objects : Layered depth prediction from a single image. *Pattern Recognition Letters*, 125:333–340, 2019. ISSN 0167-8655.

Mikael Drugge, Marcus Nilsson, Roland Parviainen et Peter Parnes : Experiences of using wearable computers for ambient telepresence and remote interaction. *In Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*, ETP '04, pages 2–11. Association for Computing Machinery, 2004. ISBN 978-1-58113-933-4.

Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, Shahram Izadi, Adarsh Kowdle, Konstantine Tsotsos et David Kim : DepthLab : Real-time 3D Interaction with Depth Maps for Mobile Augmented Reality. *In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pages 829–843. Association for Computing Machinery, 2020. ISBN 978-1-4503-7514-6.

Grégoire Dupont de Dinechin : *Towards comfortable virtual reality viewing of virtual environments created from photographs of the real world.* Thèse de doctorat, Université Paris sciences et lettres, 2020.

Grégoire Dupont de Dinechin et Alexis Paljic : Cinematic Virtual Reality With Motion Parallax From a Single Monoscopic Omnidirectional Image. *In 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, pages 1–8, 2018.

Grégoire Dupont Dupont de Dinechin et Alexis Paljic : Virtual Agents from 360° Video for Interactive Virtual Reality. *In Proceedings of the 32nd International Conference on Computer Animation and Social Agents*, CASA '19, pages 75–78. Association for Computing Machinery, 2019. ISBN 978-1-4503-7159-9.

Adrian Dziembowski, Adam Grzelka, Dawid Mieloch, Olgierd Stankiewicz, Krzysztof Wegner et Domański Domański : Multiview synthesis - Improved view synthesis for virtual navigation. *2016 Picture Coding Symposium (PCS)*, 2016.

Marc Eder, Mykhailo Shvets, John Lim et Jan-Michael Frahm : Tangent Images for Mitigating Spherical Distortion. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020.

Siavash Eftekharifar, Anne Thaler et Nikolaus Troje : Contribution of Motion Parallax and Stereopsis to the Sense of Presence in Virtual Reality. *Journal of Perceptual Imaging*, 3, 2020.

Martin Enders et Nadja Hossbach : Dimensions of Digital Twin Applications - A Literature Review. *AMCIS 2019 Proceedings*, 2019.

Douglas C. Engelbart et William K. English : A research center for augmenting human intellect. *In Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, AFIPS '68 (Fall, part I), pages 395–410. Association for Computing Machinery, 1968. ISBN 978-1-4503-7899-4.

George Fahim, Khalid Amin et Sameh Zarif : Single-View 3D reconstruction : A Survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021. ISSN 0097-8493.

Ching-Ling Fan, Wen-Chih Lo, Yu-Tung Pai et Cheng-Hsin Hsu : A Survey on 360° Video Streaming : Acquisition, Transmission, and Display. *ACM Computing Surveys*, 52(4):71 :1–71 :36, 2019. ISSN 0360-0300.

Hannes Fassold : Adapting Computer Vision Algorithms for Omnidirectional Video. *ACM Multimedia*, 2019.

Åsa Fast-Berglund, Liang Gong et Dan Li : Testing and validating Extended Reality (xR) technologies in manufacturing. *Procedia Manufacturing*, 25:31–38, 2018. ISSN 2351-9789.

Kenneth Fleischmann et Thomas Templeton : Past futures and technoscientific innovation : The mutual shaping of science fiction and science fact. *Proceedings of the American Society for Information Science and Technology*, 45:1–11, 2009.

Elodie Fourquet, William Cowan et Stephen Mann : On the empirical limits of billboard rotation. *In Proceedings of the 4th symposium on Applied perception in graphics and visualization*, APGV '07, pages 49–56. Association for Computing Machinery, 2007. ISBN 978-1-59593-670-7.

Lars Fritsche, Felix Unverzag, Jan Peters et Roberto Calandra : First-person tele-operation of a humanoid robot. *In 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 997–1002. IEEE, 2015. ISBN 978-1-4799-6885-5.

Henry Fuchs, Andrei State et Jean-Charles Bazin : Immersive 3D Telepresence. *Computer*, 47:46–52, 2014.

Philippe Fuchs : *Théorie de la réalité virtuelle : Les véritables usages*. ECOLE DES MINES, 2018. ISBN 978-2-35671-511-1.

Guillaume Gamelin, Amine Chellali, Samia Cheikh, Aylen Ricca, Cedric Dumas et Samir Otmane : Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal and Ubiquitous Computing*, 25 (3):467–484, 2021. ISSN 1617-4917.

Alejandro Gandsas, Trevor Dorey et Adrian Park : Immersive Live Streaming of Surgery Using 360-Degree Video to Head-mounted Virtual Reality Devices : A New Paradigm in Surgical Education. *Surgical Innovation*, 30(4):486–492, 2023. ISSN 1553-3514.

Lei Gao, Huidong Bai, Rob Lindeman et Mark Billinghurst : Static local environment capturing and sharing for MR remote collaboration. *In SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, SA '17, pages 1–6. Association for Computing Machinery, 2017. ISBN 978-1-4503-5410-3.

Marcel Germann, Alexander Hornung, Richard Keiser, Remo Ziegler, Stephan Würmlin et Markus Gross : Articulated Billboards for Video-based Rendering. *Computer Graphics Forum*, 29(2):585–594, 2010. ISSN 1467-8659.

Simon J. Gibbs, Constantin Arapis et Christian J. Breiteneder : TELEPORT - Towards immersive copresence. *Multimedia Systems*, 7(3):214–221, 1999. ISSN 1432-1882.

Vasileios Gkitsas, Vladimiros Sterzentsenko, Nikolaos Zioulis, Georgios Albanis et Dimitrios Zarpalas : PanoDR : Spherical Panorama Diminished Reality for Indoor Scenes. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3711–3721, 2021.

Duke Gledhill, Gui Yun Tian, Dave Taylor et David Clarke : Panoramic imaging-a review. *Computers & Graphics*, 27(3):435–445, 2003. ISSN 0097-8493.

Daniel Gotsch, Xujing Zhang, Timothy Merritt et Roel Vertegaal : TeleHuman2 : A Cylindrical Light Field Teleconferencing System for Life-size 3D Human Telepresence. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–10. Association for Computing Machinery, 2018. ISBN 978-1-4503-5620-6.

Karl Granström, Marcus Baum et Stephan Reuter : Extended Object Tracking : Introduction, Overview, and Applications. *Journal of Advances in Information Fusion*, 12, 2017.

Danillo Graziosi, Ohji Nakagami, S. Kuma, Alexandre Zaghetto, T. Suzuki et Ali Tabatabai : An overview of ongoing point cloud compression standardization activities : video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing*, 9, 2020. ISSN 2048-7703.

Markus Gross, Stephan Würmlin, Martin Naef, Edouard Lamboray, Christian Spagno, Andreas Kunz, Esther Koller-Meier, Tomas Svoboda, Luc Van Gool,

Silke Lang, Kai Strehlke, Andrew Vande Moere et Oliver Staadt : blue-c : A Spatially Immersive Display and 3D Video Portal for Telepresence. *ACM Transactions on Graphics*, 22(3):819–827, 2003. ISSN 0730-0301.

Alice Gruber et Regina Kaplan-Rakowski : Verbal and nonverbal communication in high-immersion virtual reality for language learners. *In Intelligent CALL, granular systems and learner data : short papers from EUROCALL 2022*, 2022. ISBN 978-2-38372-015-7.

Surabhi Gupta, Ashwath Shetty et Avinash Sharma : Attention based Occlusion Removal for Hybrid Telepresence Systems. *In 2022 19th Conference on Robots and Vision (CRV)*, pages 167–174, 2022.

Luis E. Gurrieri et Eric Dubois : Acquisition of omnidirectional stereoscopic images and videos of dynamic scenes : a review. *Journal of Electronic Imaging*, 22(3), 2013. ISSN 1017-9909, 1560-229X.

J. Hamill, R. McDonnell, S. Dobbyn et C. O'Sullivan : Perceptual Evaluation of Impostor Representations for Virtual Humans and Buildings. *Computer Graphics Forum*, 24(3):623–633, 2005. ISSN 1467-8659.

Hyun-Tae Han, Yoshimune Nonomura et Yuichi Tsumaki : Communication capability of telepresence system with the miniature humanoid MH-2. *Artificial Life and Robotics*, 23(3):328–337, 2018. ISSN 1614-7456.

Seo Han et Doug Suh : A 360-degree Panoramic Image Inpainting Network Using a Cube Map. *Computers, Materials & Continua*, 66:213–228, 2020.

Yuxuan Han, Ruicheng Wang et Jiaolong Yang : Single-View View Synthesis in the Wild with Learned Adaptive Multiplane Images. *In ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, pages 1–8. Association for Computing Machinery, 2022.

Takayuki Hara et Tatsuya Harada : Enhancement of Novel View Synthesis Using Omnidirectional Image Completion. *arXiv*, 2022.

Cuan M. Harrington, Dara O. Kavanagh, Gemma Wright Ballester, Athena Wright Ballester, Patrick Dicker, Oscar Traynor, Arnold Hill et Sean Tierney : 360° Operative Videos : A randomised Cross-Over Study Evaluating Attentiveness and Information retention. *Journal of Surgical Education*, 75(4):993–1000, 2018. ISSN 1931-7204.

Sandra G. Hart et Lowell E. Staveland : *Development of NASA-TLX (Task Load Index) : Results of Empirical and Theoretical Research*, volume 52 de *Human Mental Workload*, pages 139–183. North-Holland, 1988.

J. Hauber, H. Regenbrecht, A. Hills, A. Cockburn et Mark Billinghurst : Social presence in two- and three-dimensional videoconferencing. *In The 8th Annual International Workshop on Presence*, pages 189–198. University of Canterbury. Computer Science and Software Engineering., 2005.

Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du et Ken Perlin : GazeChat : Enhancing Virtual Conferences with Gaze-aware 3D Photos. *In The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 769–782. Association for Computing Machinery, 2021. ISBN 978-1-4503-8635-7.

Heiko Hecht, Evgenia Boyarskaya et Akiyoshi Kitaoka : The Mona Lisa effect : Testing the limits of perceptual robustness vis-à-vis slanted images. *Psihologija*, 47:287–301, 2014.

Peter Hedman, Suhib Alsisan, R. Szeliski et J. Kopf : Casual 3D photography. *ACM Trans. Graph.*, 2017.

Peter Hedman, Tobias Ritschel, George Drettakis et Gabriel Brostow : Scalable inside-out image-based rendering. *ACM Transactions on Graphics*, 35(6):231 :1–231 :11, 2016. ISSN 0730-0301.

Carrie Heeter : Being There : The Subjective Experience of Presence. *Presence : Teleoperators and Virtual Environments*, 1:262, 1992.

P. Heidicker, E. Langbehn et F. Steinicke : Influence of avatar appearance on presence in social VR. *In 2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 233–234, 2017.

Richard M. Held et Nathaniel I. Durlach : Telepresence. *Presence : Teleoperators and Virtual Environments*, 1(1):109–112, 1992.

Matthias Hernandez, Jongmoo Choi et Gérard Medioni : Near laser-scan quality 3-D face reconstruction from a low-quality depth stream. *Image and Vision Computing*, 36:61–69, 2015. ISSN 0262-8856.

Fernanda Herrera, Soo Youn Oh et Jeremy N. Bailenson : Effect of Behavioral Realism on Social Interactions Inside Collaborative Virtual Environments. *PRESENCE : Virtual and Augmented Reality*, 27(2):163–182, 2020. ISSN 1054-7460.

Yasamin Heshmat, Brennan Jones, Xiaoxuan Xiong, Carman Neustaedter, Anthony Tang, Bernhard E. Riecke et Lillian Yang : Geocaching with a Beam : Shared Outdoor Activities through a Telepresence Robot with 360 Degree Viewing. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13. Association for Computing Machinery, 2018. ISBN 978-1-4503-5620-6.

Juan Luis Higuera-Trujillo, Juan López-Tarruella Maldonado et Carmen Llinares Millán : Psychological and physiological human responses to simulated and real environments : A comparison between Photographs, 360° Panoramas, and Virtual Reality. *Applied Ergonomics*, 65:398–409, 2017. ISSN 0003-6870.

Deborah Hix et H Rex Hartson : *Developing user interfaces : ensuring usability through product & process*. John Wiley & Sons, Inc., 1993.

Gail Hopkins et Steve Benford : Vivid : A Symbiosis between Virtual Reality and Video Conferencing. *In UKERNA Video Conferencing Workshop*, 1996.

Ryosuke Hori, Ryo Hachiuma, Mariko Isogawa, Dan Mikami et Hideo Saito : Silhouette-Based 3D Human Pose Estimation Using a Single Wrist-Mounted 360° Camera. *IEEE Access*, 10:54957–54968, 2022. ISSN 2169-3536.

Youichi Horry, Ken-Ichi Anjyo et Kiyoshi Arai : Tour into the picture : using a spidery mesh interface to make animation from a single image. *In Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '97, pages 225–232. ACM Press/Addison-Wesley Publishing Co., 1997. ISBN 978-0-89791-896-1.

Ching-Yu Hsu, Cheng Sun et Hwann-Tzong Chen : Moving in a 360 World : Synthesizing Panoramic Parallaxes from a Single Panorama. *arXiv*, 2021.

Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker et Christian Theobalt : EgoRenderer : Rendering Human Avatars from Egocentric Camera Images. *In 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14508–14518, 2021.

Jonatan Hvass, Oliver Larsen, Kasper Vendelbo, Niels Nilsson, Rolf Nordahl et Stefania Serafin : Visual realism and presence in a virtual reality game. *In 2017 3DTV Conference : The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4, 2017.

Wijnand A. Ijsselsteijn : *History of Telepresence*, pages 5–21. John Wiley & Sons, Ltd, 2005. ISBN 978-0-470-02273-3.

Andrew Irlitti, Mesut Latifoglu, Qiushi Zhou, Martin N Reinoso, Thuong Hoang, Eduardo Velloso et Frank Vetere : Volumetric Mixed Reality Telepresence for Real-time Cross Modality Collaboration. *In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–14. Association for Computing Machinery, 2023. ISBN 978-1-4503-9421-5.

H. Ishiguro, M. Yamamoto et S. Tsuji : Omni-directional stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):257–262, 1992. ISSN 1939-3539.

Frantisek Jakab, Miroslav Michalko, Jan Turňa, Lubomir Bilský, Jana Kováčová et Dávid Cymbalák : The experience from implementation of National telepresence infrastructure in Slovakia to support research, development and technology transfer. *In 2016 International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 127–132, 2016.

Boyi Jiang, Yang Hong, Hujun Bao et Juyong Zhang : SelfRecon : Self Reconstruction Your Digital Avatar from Monocular Video. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5595–5605, 2022.

Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong et Rui Huang : UniFuse : Unidirectional Fusion for 360° Panorama Depth Estimation. *IEEE Robotics and Automation Letters*, 6:1519–1526, 2021. ISSN 2377-3766.

Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee et Mark Billinghurst : *eyemR-Vis : Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration*, pages 1–7. Association for Computing Machinery, 2021. ISBN 978-1-4503-8095-9.

Dongsik Jo, Ki-Hong Kim et Gerard Jounghyun Kim : Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems. *In Proceedings the 30th Conference on Computer Animation and Social Agents*, pages 27–36, 2017.

Robert Johansen : Current user approaches to groupware. *Groupware : Computer support for business teams*, 1988.

Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas et Paul Debevec : Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Transactions on Graphics*, 28(33):1–8, 2009. ISSN 0730-0301, 1557-7368.

Brennan Jones, Yaying Zhang, Priscilla Wong et Sean Rintel : VROOM : Virtual

Robot Overlay for Online Meetings. *In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–10. Association for Computing Machinery, 2020a. ISBN 978-1-4503-6819-3.

David Jones, Chris Snider, Aydin Nassehi, Jason Yon et Ben Hicks : Characterising the Digital Twin : A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020b. ISSN 1755-5817.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara et Yaser Sheikh : Panoptic Studio : A Massively Multiview System for Social Motion Capture. *In 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015.

Amin Jourabloo, Fernando De La Torre, Jason Saragih, Shih-En Wei, Stephen Lombardi, Te-Li Wang, Danielle Belko, Autumn Trimble et Hernan Badino : Robust Egocentric Photo-realistic Facial Expression Transfer for Virtual Reality. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20291–20300. IEEE, 2022. ISBN 978-1-66546-946-3.

Joel Jung, Bart Kroon, Renaud Doré, Gauthier Lafruit et Jill Boyce : Update on N17618 v2 CTC on 3DoF+ and Windowed 6DoF. *In 123rd MPEG meeting of ISO/IEC JTC1/SC29/WG11, MPEG123/m43571*, 2018.

Claudia Kamcke et Rainer Hutterer : *History of Dioramas*, pages 7–21. Springer Netherlands, 2015. ISBN 978-94-017-9496-1.

Takeo Kanade, Peter Rander et P.J. Narayanan : Virtualized reality : constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, 1997. ISSN 1941-0166.

Lai Kang, Jie Jiang, Yingmei Wei et Yuxiang Xie : Efficient Randomized Hierarchy Construction for Interactive Visualization of Large Scale Point Clouds. *In 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pages 593–597, 2019.

Shunichi Kasahara et Jun Rekimoto : JackIn Head : immersive visual telepresence system with omnidirectional wearable camera for remote collaboration. *In Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, VRST '15, pages 217–225. Association for Computing Machinery, 2015. ISBN 978-1-4503-3990-2.

Tadakazu Kashiwabara, Hirotaka Osawa, Kazuhiko Shinozawa et Michita Imai : TEROOS : a wearable avatar to enhance joint activities. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2001–2004. Association for Computing Machinery, 2012. ISBN 978-1-4503-1015-4.

Peter Kauff et Oliver Schreer : An immersive 3D video-conferencing system using shared virtual team user environments. *In Proceedings of the 4th international conference on Collaborative virtual environments*, pages 105–112. ACM, 2002. ISBN 978-1-58113-489-6.

Yuki Kawana, Yusuke Mukuta et Tatsuya Harada : Neural star domain as primitive representation. *In Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 7875–7886. Curran Associates Inc., 2020. ISBN 978-1-71382-954-6.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler et George Drettakis : 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Nawel Khenak, Jeanne Vézien et Patrick Bourdot : Spatial Presence, Performance, and Behavior between Real, Remote, and Virtual Immersive E]nvironments. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3467–3478, 2020. ISSN 1941-0506.

Jinwook Kim, Dooyoung Kim, Bowon Kim, Hyunchul Kim et Jeongmi Lee : Holobot : Hologram based Extended Reality Telepresence Robot. *In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 60–64. Association for Computing Machinery, 2023. ISBN 978-1-4503-9970-8.

S. Kim, H. Hagh-Shenas et V. Interrante : Conveying shape with texture : experimental investigations of texture's effects on shape categorization judgments. *IEEE Transactions on Visualization and Computer Graphics*, 10(4):471–483, 2004. ISSN 1941-0506.

Seungwon Kim, Mark Billinghurst et Gun Lee : The Effect of Collaboration Styles and View Independence on Video-Mediated Remote Collaboration. *Computer Supported Cooperative Work (CSCW)*, 27(3):569–607, 2018. ISSN 1573-7551.

Don Kimber, Patrick Proppe, Sven Kratz, Jim Vaughan, Bee Liew, Don Severns et Weiqing Su : Polly : Telepresence from a Guide's Shoulder. *In* Lourdes Agapito, Michael M. Bronstein et Carsten Rother, éditeurs : *Computer Vision - ECCV 2014 Workshops*, Lecture Notes in Computer Science, pages 509–523. Springer International Publishing, 2015. ISBN 978-3-319-16199-0.

Kazuki Kobayashi, Takashi Komuro, Keiichiro Kagawa et Shoji Kawahito : Transmission of correct gaze direction in video conferencing using screen-embedded cameras. *Multimedia Tools and Applications*, 80(21):31509–31526, 2021. ISSN 1573-7721.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge et Peter Anderson : Pathdreamer : A World Model for Indoor Navigation. *In 2021 IEEE International Conference on Computer Vision (ICCV)*, pages 14718–14728. IEEE Computer Society, 2021. ISBN 978-1-66542-812-5.

Jan Kolkmeier, Emiel Harmsen, Sander Giesselink, Dennis Reidsma, Mariet Theune et Dirk Heylen : With a little help from a holographic friend : the OpenIMPRESS mixed reality telepresence toolkit for remote collaboration systems. *In Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, pages 1–11. Association for Computing Machinery, 2018. ISBN 978-1-4503-6086-9.

Ryohei Komiyama, Takashi Miyaki et Jun Rekimoto : JackIn Space : designing a seamless transition between first and third person view for effective telepresence collaborations. *In Proceedings of the 8th Augmented Human International Conference*, AH '17, pages 1–9. Association for Computing Machinery, 2017. ISBN 978-1-4503-4835-5.

Robert Konrad, Donald G. Dansereau, Aniq Masood et Gordon Wetzstein : SpinVR : towards live-streaming 3D virtual reality video. *ACM Transactions on Graphics*, 36(6):209 :1–209 :12, 2017. ISSN 0730-0301.

Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, Peizhao Zhang, Zijian He, Peter Vajda, Ayush Saraf et Michael Cohen : One shot 3D photography. *ACM Transactions on Graphics*, 39(4):76 :76 :1–76 :76 :13, 2020. ISSN 0730-0301.

Marek Kowalski, Jacek Naruniec et Michal Daniluk : Livescan3D : A Fast and Inexpensive 3D Data Acquisition System for Multiple Kinect v2 Sensors. *In 2015 International Conference on 3D Vision*, pages 318–325, 2015.

Sven Kratz, Don Kimber, Weiqing Su, Gwen Gordon et Don Severns : Polly : "being there" through the parrot and a guide. *In Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, MobileHCI '14, pages 625–630. Association for Computing Machinery, 2014. ISBN 978-1-4503-3004-6.

Annica Kristoffersson, Silvia Coradeschi et Amy Loutfi : A Review of Mobile Robotic Telepresence. *Advances in Human-Computer Interaction*, 2013:1–17, 2013.

Shreyas Kulkarni, Peng Yin et Sebastian Scherer : 360FusionNeRF : Panoramic Neural Radiance Fields with Joint Guidance. *arXiv*, 2022.

Kartikaeya Kumar, Lev Poretski, Jiannan Li et Anthony Tang : Tourgether360 : Collaborative Exploration of 360° Videos using Pseudo-Spatial Navigation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):546 :1–546 :27, 2022.

Gregorij Kurillo et Ruzena Bajcsy : 3D teleimmersion for collaboration and interaction of geographically distributed users. *Virtual Reality*, 17(11):29–43, 2013. ISSN 1434-9957.

Kana Kushida et Hideyuki Nakanishi : PopObject : A Robotic Screen for Embodying Video-Mediated Object Presentations. *In* Hironori Egi, Takaya Yuizono, Nelson Baloian, Takashi Yoshino, Satoshi Ichimura et Armanda Rodrigues, éditeurs : *Collaboration Technologies and Social Computing*, Lecture Notes in Computer Science, pages 200–212. Springer International Publishing, 2018. ISBN 978-3-319-98743-9.

Po Kong Lai, Shuang Xie, Jochen Lang et Robert Laganière : Real-Time Panoramic Depth Maps from Omni-directional Stereo Images for 6 DoF Videos in Virtual Reality. *In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 405–412, 2019.

Joseph J. LaViola : A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1):47–56, 2000. ISSN 0736-6906.

Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz et Kevin Tong : Project Starline : A high-fidelity telepresence system. *ACM Transactions on Graphics*, 40(6):242 :1–242 :16, 2021. ISSN 0730-0301.

Chuen-Chien Lee, Ali Tabatabai et Kenji Tashiro : Free viewpoint video (FVV) survey and future research direction. *APSIPA Transactions on Signal and Information Processing*, 4, 2015. ISSN 2048-7703.

G. A. Lee, T. Teo, S. Kim et Mark Billinghurst : A User Study on MR Remote Collaboration Using Live 360 Video. *In 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 153–164, 2018.

Jungjin Lee, Bumki Kim, Kyehyun Kim, Younghui Kim et Junyong Noh : Rich360 : optimized spherical representation from structured panoramic camera arrays. *ACM Transactions on Graphics*, 35(4):63 :1–63 :11, 2016. ISSN 0730-0301.

Kyungjin Lee, Juheon Yi et Youngki Lee : FarfetchFusion : Towards Fully Mobile Live 3D Telepresence Platform. *In Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '23, pages 1–15. Association for Computing Machinery, 2023. ISBN 978-1-4503-9990-6.

Kyungjin Lee, Juheon Yi, Youngki Lee, Sunghyun Choi et Young Min Kim : GROOT : A Real-time Streaming System of High-Fidelity Volumetric Videos. *In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, pages 1–14. Association for Computing Machinery, 2020. ISBN 978-1-4503-7085-1.

Yongjae Lee, Byounghyun Yoo et Soo-Hong Lee : Sharing Ambient Objects Using Real-time Point Cloud Streaming in Web-based XR Remote Collaboration. *In The 26th International Conference on 3D Web Technology*, Web3D '21, pages 1–9. Association for Computing Machinery, 2021. ISBN 978-1-4503-9095-8.

J. Leigh, A.E. Johnson, T.A. DeFanti, M. Brown, M.D. Ali, S. Bailey, A. Banerjee, P. Benerjee, Jim Chen, K. Curry, J. Curtis, F. Dech, B. Dodds, I. Foster, S. Fraser, K. Ganeshan, D. Glen, R. Grossman, R. Heiland, J. Hicks, A.D. Hudson, T. Imai, M.A. Khan, A. Kapoor, R.V. Kenyon, J. Kelso, R. Kriz, C. Lascara, X. Liu, Y. Lin, T. Mason, A. Millman, K. Nobuyuki, K. Park, B. Parod, P.J. Rajlich, M. Rasmussen, M. Rawlings, D.H. Robertson, S. Thongrong, R.J. Stein, K. Swartz, S. Tuecke, H. Wallach, Hong Yee Wong et G.H. Wheless : A review of tele-immersive applications in the cave research network. *In Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, pages 180–187, 1999.

Jason Leigh, Thomas A. DeFanti, Andrew Johnson, Maxine Brown et Daniel J. Sandin : Global Tele-Immersion : Better than Being There. *In 7th International Conference on Artificial Reality and Tele-Existence*, 1997.

Daniel Leithinger, Sean Follmer, Alex Olwal et Hiroshi Ishii : Physical telepresence : shape capture and display for embodied, computer-mediated remote collaboration. *In Proceedings of the 27th annual ACM symposium on User interface software and technology*, UIST '14, pages 461–470. Association for Computing Machinery, 2014. ISBN 978-1-4503-3069-5.

Laurent Lescop : 360° vision, from panoramas to VR. *In Envisioning Architecture SPACE / TIME / MEANING*, volume 1. Mackintosh School of Architecture and the School of Simulation and Visualization at the Glasgow School of Art, 2017.

Lin Li, Salman Khan et Nick Barnes : Silhouette-Assisted 3D Object Instance Reconstruction from a Cluttered Scene. *In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2080–2088, 2019.

Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang et Hao Li : Volumetric Human Teleportation. *In ACM SIGGRAPH 2020 Real-Time Live*, SIGGRAPH 2020. Association for Computing Machinery, 2020. ISBN 9781450380607.

Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan et Liu Ren : OmniFusion : 360 Monocular Depth Estimation via Geometry-Aware Fusion. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2791–2800. IEEE Computer Society, 2022. ISBN 978-1-66546-946-3.

Jyh-Ming Lien : Approximate Star-Shaped Decomposition of Point Set Data. *In Eurographics Symposium on Point-Based Graphics*, pages 73–80. The Eurographics Association, 2007. ISBN 978-3-905673-51-7.

Kai-En Lin, Zexiang Xu, Ben Mildenhall, Pratul P. Srinivasan, Yannick Hold-Geoffroy, Stephen DiVerdi, Qi Sun, Kalyan Sunkavalli et Ravi Ramamoorthi : Deep Multi Depth Panoramas for View Synthesis. *In Computer Vision - ECCV 2020*, pages 328–344. Springer-Verlag, 2020. ISBN 978-3-030-58600-3.

Peter Lincoln, Greg Welch, Andrew Nashel, Adrian Ilie, Andrei State et Henry Fuchs : Animatronic Shader Lamps Avatars. *In 2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 27–33, 2009.

Salvatore Livatino : Photorealistic VR Games ? *In 17th International Conference on Artificial Reality and Telexistence (ICAT 2007)*, pages 292–293, 2007.

Shih-Yu Lo et Chih-Yuan Lai : Investigating how immersive virtual reality and active navigation mediate the experience of virtual concerts. *Scientific Reports*, 13(11):8507, 2023. ISSN 2045-2322.

Matthew Lombard et Theresa Ditton : At the Heart of It All : The Concept of Presence. *Journal of Computer-Mediated Communication*, 3(2), 1997. ISSN 1083-6101.

Andrei-Iuliu Lucaci, Morten Bach Jakobsen, Poul Anker Jensen et Claus Brøndgaard Madsen : Influence of Texture Fidelity on Spatial Perception in Virtual Reality. *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - GRAPP*, pages 244–251, 2022. ISSN 978-989-758-555-5.

Le Luo, Dongdong Weng, Jie Hao, Ziqi Tu et Haiyan Jiang : Viewpoint-Controllable Telepresence : A Robotic-Arm-Based Mixed-Reality Telecollaboration System. *Sensors*, 23(88):4113, 2023. ISSN 1424-8220.

Alberto López-Cerón et José Cañas : Accuracy analysis of marker-based 3D visual localization. *In Actas de las XXXVII Jornadas de Automática*, pages 1124–1131, 2022.

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre et Yaser Sheikh : Pixel Codec Avatars. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021.

Andrew MacQuarrie et Anthony Steed : Cinematic virtual reality : Evaluating the effect of display type on the viewing experience for panoramic video. *In 2017 IEEE Virtual Reality (VR)*, pages 45–54, 2017.

Andrew Maimone et Henry Fuchs : A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Life-Sized Tracked Display Wall. *Proceedings of ICAT 2011*, pages 4–9, 2011.

Andrew Maimone et Henry Fuchs : Real-time volumetric 3D capture of room-sized scenes for telepresence. *In 2012 3DTV-Conference : The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4, 2012.

Gordon M. Mair : How Fiction Informed the Development of Telepresence and Teleoperation. *In* Randall Shumaker, éditeur : *Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments*, pages 368–377. Springer, 2013. ISBN 978-3-642-39405-8.

Divine Maloney, Guo Freeman et Donghee Yvette Wohn : "Talking without a Voice" : Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):175 :1–175 :25, 2020.

Virginia Mamone, Fabrizio Cutolo, Sara Condino et Vincenzo Ferrari : Projected Augmented Reality to Guide Manual Precision Tasks : An Alternative to Head Mounted Displays. *IEEE Transactions on Human-Machine Systems*, pages 1–11, 2021. ISSN 2168-2305.

Steve Mann, Tom Furness, Yu Yuan, Jay Iorio et Zixin Wang : All Reality : Virtual, Augmented, Mixed (x), Mediated (x,y), and Multimediated reality. *arXiv*, 2018.

Eric Marchand, Hideaki Uchiyama et Fabien Spindler : Pose Estimation for Augmented Reality : A Hands-On Survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633, 2016.

Akira Matsuda, Kazunori Nozawa et Jun Rekimoto : JackIn Neck : A Neckband Wearable Telepresence System Designed for High Comfortability. *In Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, ISS '18, pages 415–418. Association for Computing Machinery, 2018. ISBN 978-1-4503-5694-7.

Giuseppe Mazzola, Liliana Lo Presti, Edoardo Ardizzone et Marco La Cascia : A Dataset of Annotated Omnidirectional Videos for Distancing Applications. *Journal of Imaging*, 7(8):158, 2021. ISSN 2313-433X.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu et Christian Theobalt : Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. *In 2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017.

Andreas Meuleman, Hyeonjoong Jang, Daniel S. Jeon et Min H. Kim : Real-Time Sphere Sweeping Stereo from Multiview Fisheye Images. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11418–11427. IEEE, 2021. ISBN 978-1-66544-509-2.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi et Ren Ng : Nerf : representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. ISSN 0001-0782.

Paul Milgram, Haruo Takemura, Akira Utsumi et Fumio Kishino : Augmented reality : a class of displays on the reality-virtuality continuum. *In Telemanipulator and Telepresence Technologies*, volume 2351, pages 282–292. SPIE, 1995.

Aleksandar Minja et Vojin Šenk : Quasi-Analytical Simulation Method for Estimating the Error Probability of Star Domain Decoders. *IEEE Transactions on Communications*, 67(5):3101–3113, 2019. ISSN 1558-0857.

Marvin Minsky : Telepresence. *Omni*, 2(9):44–52, 1980.

Kana Misawa, Yoshio Ishiguro et Jun Rekimoto : LiveMask : a telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. *In Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 394–397. Association for Computing Machinery, 2012a. ISBN 978-1-4503-1287-5.

Kana Misawa, Yoshio Ishiguro et Jun Rekimoto : Ma petite chérie : What are you looking at ? A Small Telepresence System to Support Remote Collaborative Work for Intimate Communication. *In Proceedings of the 3rd augmented human international conference*, pages 1–5, 2012b.

Kana Misawa et Jun Rekimoto : Wearing another's personality : a human-surrogate system with a telepresence face. *In Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ISWC '15, pages 125–132. Association for Computing Machinery, 2015. ISBN 978-1-4503-3578-2.

Tom Monnier, Matthew Fisher, Alexei A. Efros et Mathieu Aubry : Share With Thy Neighbors : Single-View Reconstruction by Cross-Instance Consistency. *arXiv*, 2022.

Vito Monteleone, Liliana Lo Presti et Marco La Cascia : Pedestrian Tracking in 360 Video by Virtual PTZ Cameras. *In 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI)*, pages 1–6, 2018.

Masahiro Mori, Karl F. MacDorman et Norri Kageki : The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. ISSN 1558-223X.

Jane Mulligan et Kostas Daniilidis : Real time trinocular stereo for tele-immersion. *In Proceedings 2001 International Conference on Image Processing*, volume 3, pages 959–962, 2001.

Moritz Mühlhausen, Moritz Kappel, Marc Kassubeck, Paul Bittner, Susana Castillo et Marcus Magnor : Temporal Consistent Motion Parallax for Omnidirectional Stereo Panorama Video. *In Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, pages 1–9. Association for Computing Machinery, 2020.

Arjun Nagendran, Anthony Steed, Brian Kelly et Ye Pan : Symmetric telepresence using robotic humanoid surrogates. *Computer Animation and Virtual Worlds*, 26(3-4):271–280, 2015. ISSN 1546-427X.

David Narciso, Maximino Bessa, Miguel Melo, António Coelho et José Vasconcelos-Raposo : Immersive 360° video user experience : impact of different variables in the sense of presence and cybersickness. *Universal Access in the Information Society*, 18(1):77–87, 2019. ISSN 1615-5297.

Erica E. Nason, Mark Trahan, Scott Smith, Vangelis Metsis et Katherine Selber : Virtual treatment for veteran social anxiety disorder : A comparison of 360° video and

3D virtual reality. *Journal of Technology in Human Services*, 38(3):288–308, 2020. ISSN 1522-8835.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges et Andrew Fitzgibbon : KinectFusion : Real-time dense surface mapping and tracking. *In 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann et Feng Liu : CollaVR : Collaborative In-Headset Review for VR Video. *In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pages 267–277. Association for Computing Machinery, 2017. ISBN 978-1-4503-4981-9.

Matthias Niessner, Michael Zollhöfer, Shahram Izadi et Marc Stamminger : Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics*, 32(6):169 :1–169 :11, 2013. ISSN 0730-0301.

Toshikazu Nishimura, Hideyuki Nakanishi, Chikara Yoshida et Toru Ishida : Applying videogame technologies to video conferencing systems. *In Proceedings of the 1998 ACM symposium on Applied Computing*, SAC '98, pages 471–476. Association for Computing Machinery, 1998. ISBN 978-0-89791-969-2.

Shuichi Nishio, Hiroshi Ishiguro et Norihiro Hagita : Geminoid : Teleoperated android of an existing person. *Humanoid robots : New developments*, 14(343-352):10–1109, 2007.

Kohei Ogawa, Shuichi Nishio, Kensuke Koda, Giuseppe Balistreri, Tetsuya Watanabe et Hiroshi Ishiguro : Exploring the Natural Reaction of Young and Aged Person with Telenoid in a Real World. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 15:592–597, 2011.

Tetsurou Ogi et Yasuto Fueki : Development of head mounted display based tele-immersion system for collaborative work. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(3-2):21–25, 2017. ISSN 2180-1843.

Catherine S. Oh, Jeremy N. Bailenson et Gregory F. Welch : A Systematic Review of Social Presence : Definition, Antecedents, and Implications. *Frontiers in Robotics and AI*, 5, 2018. ISSN 2296-9144.

Stephan Ohl : Tele-Immersion Concepts. *IEEE Transactions on Visualization and Computer Graphics*, 24(10):2827–2842, 2018. ISSN 1941-0506.

Judith S. Olson et Gary M. Olson : *Computer-Supported Cooperative Work*, pages 243–253. Elsevier, 2003. ISBN 978-0-12-227240-0.

Yuya Onishi, Kazuaki Tanaka et Hideyuki Nakanishi : Embodiment of Video-mediated Communication Enhances Social Telepresence. *In Proceedings of the Fourth International Conference on Human Agent Interaction*, HAI '16, pages 171–178. Association for Computing Machinery, 2016. ISBN 978-1-4503-4508-8.

Marta Orduna, Jesús Gutiérrez, Alejandro Sánchez, Julián Cabrera, César Díaz, Pablo Perez et Narciso García : Evaluation of the Performance of an Immersive System for Tele-education. *In ACM International Conference on Interactive Media Experiences*, pages 209–220. ACM, 2022. ISBN 978-1-4503-9212-9.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin et Shahram Izadi : Holoportation : Virtual 3D Teleportation in Real-time. *In Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754. Association for Computing Machinery, 2016. ISBN 978-1-4503-4189-9.

Kazuhiro Otsuka : MMSpace : Kinetically-augmented telepresence for small group-to-group conversations. *In 2016 IEEE Virtual Reality (VR)*, pages 19–28, 2016.

Mai Otsuki, Taiki Kawano, Keita Maruyama, Hideaki Kuzuoka et Yusuke Suzuki : ThirdEye : Simple Add-on Display to Represent Remote Participant's Gaze Direction in Video Communication. *In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5307–5312. Association for Computing Machinery, 2017. ISBN 978-1-4503-4655-9.

Ryan S. Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr et Paul Debevec : A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics*, 37(6):197 :1–197 :15, 2018. ISSN 0730-0301.

Alper Ozkan et Ufuk Celikcan : The relationship between cybersickness and eye-activity in response to varying speed, scene complexity and stereoscopic vr parameters. *International Journal of Human-Computer Studies*, 176:103039, 2023. ISSN 1071-5819.

Ye Pan et Anthony Steed : A Comparison of Avatar-, Video-, and Robot-Mediated Interaction on Users' Trust in Expertise. *Frontiers in Robotics and AI*, 3, 2016. ISSN 2296-9144.

Eric Paulos et John Canny : Ubiquitous tele-embodiment : applications and implications. *International Journal of Human-Computer Studies*, 46(6):861–877, 1997. ISSN 1071-5819.

Eric Paulos et John Canny : Social Tele-embodiment : Understanding Presence. *Autonomous Robots*, 11(1):87–95, 2001. ISSN 1573-7527.

Fabrizio Pece, William Steptoe, Fabian Wanner, Simon Julier, Tim Weyrich, Jan Kautz et Anthony Steed : Panoinserts : mobile spatial teleconferencing. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1319–1328. Association for Computing Machinery, 2013. ISBN 978-1-4503-1899-0.

Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek et Andrew Wilson : Room2Room : Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. *In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1716–1725. Association for Computing Machinery, 2016. ISBN 978-1-4503-3592-8.

R. L. Pepper et J. D. Hightower : Research Issues in Teleoperator Systems. *Proceedings of the Human Factors Society Annual Meeting*, 28(9):803–807, 1984. ISSN 0163-5182.

Kevin Pfeil, Pamela Wisniewski et Joseph J. LaViola Jr. : An Analysis of User Perception Regarding Body-Worn 360° Camera Placements and Heights for Telepresence. *In ACM Symposium on Applied Perception 2019*, SAP '19, pages 1–10. Association for Computing Machinery, 2019. ISBN 978-1-4503-6890-2.

Giovanni Pintore, Marco Agus, Eva Almansa et Enrico Gobbetti : Instant Automatic Emptying of Panoramic Indoor Scenes. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3629–3639, 2022. ISSN 1941-0506.

Thammathip Piumsomboon, Gun A. Lee, Andrew Irlitti, Barrett Ens, Bruce H. Thomas et Mark Billinghurst : On the Shoulder of the Giant : A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–17. Association for Computing Machinery, 2019. ISBN 978-1-4503-5970-2.

Claudia Plüss, Nicola Ranieri, Jean-Charles Bazin, Tobias Martin, Pierre-Yves Laffont, Tiberiu Popa et Markus Gross : An Immersive Bidirectional System for Life-size 3D Communication. *In Proceedings of the 29th International Conference on Computer Animation and Social Agents*, CASA '16, pages 89–96. Association for Computing Machinery, 2016. ISBN 978-1-4503-4745-7.

Ronald Poelman, Oytun Akman, Stephan Lukosch et Pieter Jonker : As if being there : Mediated reality for crime scene investigation. *In Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 1267–1276, 2012.

David Raitt, P Gyger et A Woods : Innovative technologies from science fiction for space applications. *Preparing for the Future*, 11(1):6–7, 2001.

Thinal Raj, Fazida Hanim Hashim, Aqilah Baseri Huddin, Mohd Faisal Ibrahim et Aini Hussain : A Survey on LiDAR Scanning Mechanisms. *Electronics*, 9(55):741, 2020. ISSN 2079-9292.

Rene Ranftl, Katrin Lasinger, David Hafner et Vladlen Koltun : Towards Robust Monocular Depth Estimation : Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin et Henry Fuchs : The Office of the Future : A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998*, 1998.

Majken K. Rasmussen, Esben W. Pedersen, Marianne G. Petersen et Kasper Hornbæk : Shape-changing interfaces : a review of the design space and open research questions. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 735–744. Association for Computing Machinery, 2012. ISBN 978-1-4503-1015-4.

Philipp A. Rauschnabel, Reto Felix, Chris Hinsch, Hamza Shahab et Florian Alt : What is XR ? Towards a Framework for Augmented and Virtual Reality. *Computers in Human Behavior*, 133:107289, 2022. ISSN 0747-5632.

Manuel Rey-Area, Mingze Yuan et Christian Richardt : 360MonoDepth : High-Resolution 360° Monocular Depth Estimation. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3752–3762. IEEE Computer Society, 2022. ISBN 978-1-66546-946-3.

Taehyun Rhee, Stephen Thompson, Daniel Medeiros, Rafael Anjos et Andrew Chalmers : Augmented Virtual Teleportation for High-Fidelity Telecollaboration. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 2020.

Christian Richardt : Omnidirectional Stereo. *Computer Vision : A Reference Guide*, 2020.

Christian Richardt, James Tompkin et Gordon Wetzstein : *Capture, Reconstruction, and Representation of the Visual Real World for Virtual Reality*, pages 3–32. Lecture Notes in Computer Science. Springer International Publishing, 2020. ISBN 978-3-030-41816-8.

K. A. Ritter et Terrence L. Chambers : Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training. *Virtual Reality*, 2021. ISSN 1434-9957.

David J. Roberts, Allen J. Fairchild, Simon P. Campion, John O'Hare, Carl M. Moore, Rob Aspin, Tobias Duckworth, Paolo Gasparello et Franco Tecchia : withyou - An experimental end-to-end Telepresence System Using Video-Based Reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 9(3):562–574, 2015. ISSN 1941-0484.

Jeffrey J. Roth, Mari Pierce et Steven Brewer : Performance and Satisfaction of Resident and Distance Students in Videoconference Courses. *Journal of Criminal Justice Education*, 31(2):296–310, 2020. ISSN 1051-1253.

Amela Sadagic, H. Towles, J. Lanier, H. Fuchs, Andries van Dam, Kostas Daniilidis, Jane Mulligan, Loring Holden et Bob Zeleznik : National tele-immersion initiative : Towards compelling tele-immersive collaborative environments. Presentation given at Medicine meets Virtual Reality 2001 Conference, 2001.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa et Hao Li : PIFu : Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. *In The IEEE International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih et Hanbyul Joo : PIFuHD : Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. ISBN 978-1-72817-168-5.

Daisuke Sakamoto, Takayuki Kanda, Tetsuo Ono, Hiroshi Ishiguro et Norihiro Hagita : Android as a Telecommunication Medium with a Human-like Presence. *In Proceedings of the ACM/IEEE international conference on Human-robot interaction*, HRI '07, pages 193–200. Association for Computing Machinery, 2007. ISBN 978-1-59593-617-2.

Mose Sakashita, E. Andy Ricci, Jatin Arora et François Guimbretière : RemoteCoDe : Robotic Embodiment for Enhancing Peripheral Awareness in Remote Collaboration Tasks. *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW1):63 :1–63 :22, 2022.

N. Sakata, T. Kurata, T. Kato, M. Kourogi et H. Kuzuoka : WACL : supporting telecommunications using - wearable active camera with laser pointer. *In Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 53–56, 2003.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz et Andrew Rabinovich : SuperGlue : Learning Feature Matching with Graph Neural Networks. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2020. ISBN 978-1-72817-168-5.

Davide Scaramuzza : Omnidirectional Camera. *Computer Vision : A Reference*, 2014.

Ryan Schubert, Greg Welch, Peter Lincoln, Arjun Nagendran, Remo Pillat et Henry Fuchs : Advances in Shader Lamps Avatars for telepresence. *In 2012 3DTV-Conference : The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4, 2012.

Sebastian Schwarz, Miska M. Hannuksela, Vida Fakour-Sevom et Nahid Sheikhi-Pour : 2D Video Coding of Volumetric Video Data. *In 2018 Picture Coding Symposium (PCS)*, pages 61–65, 2018.

Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A. Chou, Robert A. Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, Joan Llach, Khaled Mammou, Rufael Mekuria, Ohji Nakagami, Ernestasia Siahaan, Ali Tabatabai, Alexis M. Tourapis et Vladyslav Zakharchenko : Emerging MPEG Standards for Point Cloud Compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2019. ISSN 2156-3365.

Alexander Schäfer, Gerd Reis et Didier Stricker : Investigating the Sense of Presence Between Handcrafted and Panorama Based Virtual Environments. *In Mensch und Computer 2021*, MuC '21, pages 402–405. Association for Computing Machinery, 2021. ISBN 978-1-4503-8645-6.

Alexander Schäfer, Gerd Reis et Didier Stricker : A Survey on Synchronous Augmented, Virtual, and Mixed Reality Remote Collaboration Systems. *ACM Computing Surveys*, 55(66):1–27, 2023. ISSN 0360-0300, 1557-7341r.

Markus Schütz, Gottfried Mandlburger, Johannes Otepka et Michael Wimmer : Progressive Real-Time Rendering of One Billion Points Without Hierarchical Acceleration Structures. *Computer Graphics Forum*, 39(2):51–64, 2020. ISSN 1467-8659.

Derar Serhan : Transitioning from Face-to-Face to Remote Learning : Students' Attitudes and Perceptions of using Zoom during COVID-19 Pandemic. *International Journal of Technology in Education and Science*, 4(44):335–342, 2020. ISSN 2651-5369.

Ana Serrano, Incheol Kim, Zhili Chen, S. DiVerdi, D. Gutierrez, Aaron Hertzmann et B. Masiá : Motion parallax for 360° RGBD video. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

Jonathan Shade, Steven Gortler, Li-wei He et Richard Szeliski : Layered depth images. *In Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '98, pages 231–242. Association for Computing Machinery, 1998. ISBN 978-0-89791-999-9.

Rabia Shafi, Wan Shuai et Muhammad Usman Younus : 360-Degree Video Streaming : A survey of the State of the Art. *Symmetry*, 12(99), 2020.

Hanieh Shakeri et Carman Neustaedter : Teledrone : Shared Outdoor Exploration Using Telepresence Drones. *In Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 367–371, 2019. ISBN 978-1-4503-6692-2.

Matthew Shere, Hansung Kim et Adrian Hilton : 3D Human Pose Estimation From Multi Person Stereo 360 Scenes. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2019.

T. B. Sheridan : Teleoperation, telerobotics and telepresence : A progress report. *Control Engineering Practice*, 3(2):205–214, 1995. ISSN 0967-0661.

Thomas Sheridan : Musings on Telepresence and Virtual Presence. *Presence*, 1:120–125, 1992.

Meng-Li Shih, Shih-Yang Su, Johannes Kopf et Jia-Bin Huang : 3D Photography using Context-aware Layered Depth Inpainting. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. ISBN 978-1-72817-168-5.

Keisuke Shiro, Atsushi Okada, Takashi Miyaki et Jun Rekimoto : OmniGaze : A Display-covered Omnidirectional Camera for Conveying Remote User's Presence. *In Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI

'18, pages 176–183. Association for Computing Machinery, 2018. ISBN 978-1-4503-5953-5.

Harry Shum et Sing Bing Kang : A Review of Image-Based Rendering Techniques. *In Visual Communications and Image Processing*, volume 4067, pages 2–13, 2000.

Stephen Siemonsma et Tyler Bell : HoloKinect : Holographic 3D Video Conferencing. *Sensors*, 22(2121):8118, 2022. ISSN 1424-8220.

David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee et Michael Sinclair : Motion and Attention in a Kinetic Videoconferencing Proxy. *In Human-Computer Interaction - INTERACT 2011 - 13th IFIP TC 13 International Conference*, volume 6946, page 180, 2011. ISBN 978-3-642-23773-7.

Richard Skarbez, Frederick P. Brooks, Jr. et Mary C. Whitton : A Survey of Presence and Related Concepts. *ACM Computing Surveys*, 50(6):96 :1–96 :39, 2017. ISSN 0360-0300.

Mel Slater : A Note on Presence Terminology. *Presence Connect*, 3, 2003.

Mel Slater : Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 364(1535):3549–3557, 2009. ISSN 0962-8436.

Harrison Jesse Smith et Michael Neff : Communication Behavior in Embodied Virtual Reality. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12. Association for Computing Machinery, 2018. ISBN 978-1-4503-5620-6.

Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey et Giuliano Maciocci : BeThere : 3D mobile collaboration with spatial input. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 179–188. Association for Computing Machinery, 2013. ISBN 978-1-4503-1899-0.

Konstantin Sofiiuk, Ilia A. Petrov et Anton Konushin : Reviving Iterative Training with Mask Guidance for Interactive Segmentation. *In 2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145, 2022.

Youngbin Son, Seongwon Yoon, Se-Young Oh et Soohee Han : A Lightweight and Cost-Effective 3D Omnidirectional Depth Sensor Based on Laser Triangulation. *IEEE Access*, 7:58740–58750, 2019. ISSN 2169-3536.

Robert Speidel, Edward Felder, Achim Schneider et Wolfgang Öchsner : Virtual reality against Zoom fatigue ? a field study on the teaching and learning experience in interactive video and VR conferencing. *GMS Journal for Medical Education*, 40 (22), 2023. ISSN 2366-5017.

Anthony Steed, William Steptoe, Wole Oyekoya, Fabrizio Pece, Tim Weyrich, Jan Kautz, Doron Friedman, Angelika Peer, Massimiliano Solazzi, Franco Tecchia, Massimo Bergamasco et Mel Slater : Beaming : An Asymmetric Telepresence System. *Computer Graphics and Applications, IEEE*, 32:10–17, 2012.

Patrick Stotko, Stefan Krumpen, Matthias B. Hullin, Michael Weinmann et Reinhard Klein : SLAMCast : Large-Scale, Real-Time 3D Reconstruction and Streaming for Immersive Multi-Client Live Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2102–2112, 2019a. ISSN 1077-2626, 1941-0506, 2160-9306.

Patrick Stotko, Stefan Krumpen, Max Schwarz, Christian Lenz, Sven Behnke, Reinhard Klein et Michael Weinmann : A VR System for Immersive Teleoperation and Live Exploration with a Mobile Robot. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3630–3637, 2019b.

Miriam Sturdee et Jason Alexander : Analysis and Classification of Shape-Changing Interfaces for Design and Application-based Research. *ACM Computing Surveys*, 51 (1):2 :1–2 :32, 2018. ISSN 0360-0300.

Austin Sumigray, Eliot Laidlaw, James Tompkin et Stefanie Tellex : Improving Remote Environment Visualization through 360 6DoF multi-sensor fusion for VR Telerobotics. *In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, pages 387–391. Association for Computing Machinery, 2021. ISBN 978-1-4503-8290-8.

Cheng Sun, Chi-Wei Hsiao, Min Sun et Hwann-Tzong Chen : HorizonNet : Learning Room Layout with 1D Representation and Pano Stretch Data Augmentation. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056. IEEE Computer Society, 2019. ISBN 978-1-72813-293-8.

Cheng Sun, Min Sun et Hwann-Tzong Chen : HoHoNet : 360 Indoor Holistic Understanding with Latent Horizontal Features. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2021. ISBN 978-1-66544-509-2.

Haoxi Sun et Stefanie Zollmann : Towards Casually Captured 6DoF VR Videos. *In 2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 176–179, 2022.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park et Victor Lempitsky : Resolution-robust Large Mask Inpainting with Fourier Convolutions. *In 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182. IEEE, 2022. ISBN 978-1-66540-915-5.

Tim Szigeti, Kevin McMenamy, Roland Saville et Alan Glowacki : *Cisco telepresence fundamentals*. Cisco Press, 2009.

Susumu Tachi : Forty Years of Telexistence - From Concept to TELESAR VI. *ICAT-EGVE 2019 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 2019. ISSN 1727-530X.

Susumu Tachi, Yasuyuki Inoue et Fumihiro Kato : TELESAR VI : Telexistence Surrogate Anthropomorphic Robot VI. *International Journal of Humanoid Robotics*, 17(05):2050019, 2020. ISSN 0219-8436.

Susumu Tachi, Naoki Kawakami, Hideaki Nii, Kouichi Watanabe et Kouta Minamizawa : TELEsarPHONE : Mutual Telexistence Master-Slave Communication System Based on Retroreflective Projection Technology. *SICE Journal of Control, Measurement, and System Integration*, 2008.

Matthew Tait et Mark Billinghurst : The Effect of View Independence in a Collaborative AR System. *Computer Supported Cooperative Work (CSCW)*, 24(6):563–589, 2015. ISSN 1573-7551.

Zhipeng Tan, Yuning Hu et Kun Xu : Virtual Reality Based Immersive Telepresence System for Remote Conversation and Collaboration. *In* Jian Chang, Jian Jun Zhang, Nadia Magnenat Thalmann, Shi-Min Hu, Ruofeng Tong et Wencheng Wang, éditeurs : *Next Generation Computer Animation Techniques*, Lecture Notes in Computer Science, pages 234–247. Springer International Publishing, 2017. ISBN 978-3-319-69487-0.

Anthony Tang, Omid Fakourfar, Carman Neustaedter et Scott Bateman : Collaboration with 360° Videochat : Challenges and Opportunities. *In Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, pages 1327–1339. Association for Computing Machinery, 2017. ISBN 978-1-4503-4922-2.

Ravi Tejwani, Chengyuan Ma, Paolo Bonato et H. Harry Asada : An Avatar Robot Overlaid with the 3D Human Model of a Remote Operator. *In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7061–7068, 2023.

Theophilus Teo, Louise Lawrence, Gun A. Lee, Mark Billinghurst et Matt Adcock : Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14. Association for Computing Machinery, 2019. ISBN 978-1-4503-5970-2.

Ingvar Tjostheim, Wolfgang Leister et J. A. Waterworth : *Telepresence and the Role of the Senses*, pages 169–187. Philosophical Studies Series. Springer International Publishing, 2019. ISBN 978-3-030-01800-9.

Hiroaki Tobita : Gutsy-Avatar : Computational Assimilation for Advanced Communication and Collaboration. *In 2017 First IEEE International Conference on Robotic Computing (IRC)*, pages 8–13, 2017.

Hiroaki Tobita, Shigeaki Maruyama et Takuya Kuzi : Floating avatar : telepresence system using blimps for communication and entertainment. *In CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 541–550. Association for Computing Machinery, 2011. ISBN 978-1-4503-0268-5.

Yutaka Tokuda, Atsushi Hiyama, Takahiro Miura, Tomohiro Tanikawa et Michitaka Hirose : Towards Mobile Embodied 3D Avatar as Telepresence Vehicle. *In* Constantine Stephanidis et Margherita Antona, éditeurs : *Universal Access in Human-Computer Interaction. Applications and Services for Quality of Life*, Lecture Notes in Computer Science, pages 671–680. Springer, 2013. ISBN 978-3-642-39194-1.

Herman Towles, Wei-Chao Chen, Ruigang Yang, Sang-uok Kum, Henry Fuchs, Carolina Hill, Nikhil Kelshikar, Jane Mulligan, Loring Holden, Bob Brown, Amela Sadagic et Jaron Lanier : 3D Tele-Collaboration Over Internet2. *In International Workshop on Immersive Telepresence*, 2003.

Y. Tsumaki, Y. Fujita, A. Kasai, C. Sato, D.N. Nenchev et M. Uchiyama : Telecommunicator : a novel robot system for human communications. *In 11th IEEE International Workshop on Robot and Human Interactive Communication Proceedings*, pages 35–40, 2002.

Yuichi Tsumaki, Fumiaki Ono et Taisuke Tsukuda : The 20-DOF miniature humanoid MH-2 : A wearable communication system. *In 2012 IEEE International Conference on Robotics and Automation*, pages 3930–3935, 2012.

Richard Tucker et Noah Snavely : Single-View View Synthesis With Multiplane Images. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–557. IEEE, 2020. ISBN 978-1-72817-168-5.

Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat et Pablo Cesar : VR2Gather : A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia*, 30(2):48–59, 2023. ISSN 1941-0166.

John Waidhofer, Richa Gadgil, Anthony Dickson, Stefanie Zollmann et Jonathan Ventura : PanoSynthVR : View Synthesis From A Single Input Panorama with Multi-Cylinder Images. *In 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 584–592, 2022.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu et Yi-Hsuan Tsai : BiFuse : Monocular 360 Depth Estimation via Bi-Projection Fusion. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2020.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu et Yi-Hsuan Tsai : LED2-Net : Monocular 360 Layout Estimation via Differentiable Depth Rendering. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12951–12960. IEEE Computer Society, 2021a. ISBN 978-1-66544-509-2.

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar et Christian Theobalt : Estimating Egocentric 3D Human Pose in Global Space. *In 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11480–11489, 2021b.

Junying Wang, Jae Shin Yoon, Tuanfeng Y. Wang, Krishna Kumar Singh et Ulrich Neumann : Complete 3D Human Reconstruction from a single incomplete image. *In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8748–8758. IEEE, 2023a. ISBN 9798350301298.

Shengze Wang, Ziheng Wang, Ryan Schmelzle, Liujie Zheng, YoungJoong Kwon, Soumyadip Sengupta et Henry Fuchs : Bringing Telepresence to Every Desk. *arXiv*, 2023b.

Séamas Weech, Sophie Kenny et Michael Barnett-Cowan : Presence and Cybersickness in Virtual Reality Are negatively Related : A Review. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078.

Krzysztof Wegner, Olgierd Stankiewicz, Tomasz Grajek et Marek Domański : Depth map formats used within MPEG 3D frameworks. *ISO/IEC JTC1/SC29/WG11 MPEG2017 M*, 40019, 2017.

Robert B. Welch, Theodore T. Blackmon, Andrew Liu, Barbara A. Mellers et Lawrence W. Stark : The Effects of Pictorial Realism, Delay of Visual Feedback, and Observer Interactivity on the Subjective Sense of Presence. *Presence : Teleoperators and Virtual Environments*, 5(3):263–273, 1996.

Michael Wimmer et Claus Scheiblauer : Instant points : fast rendering of unprocessed point clouds. *In Proceedings of the 3rd Eurographics / IEEE VGTC conference on Point-Based Graphics*, SPBG'06, pages 129–137. Eurographics Association, 2006. ISBN 978-3-905673-32-6.

Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang et Shenghua Gao : Layout-Guided Novel View Synthesis from a Single Indoor Panorama. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16433–16442. IEEE Computer Society, 2021a. ISBN 978-1-66544-509-2.

M. Xu, C. Li, S. Zhang et P. L. Callet : State-of-the-Art in 360° Video/Image Processing : Perception, Assessment and Compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. ISSN 1941-0484.

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech et Ulrich Neumann : DISN : Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction. *arXiv*, 2021b.

Xiangyu Xu et Chen Change Loy : 3D human texture estimation from a single image with transformers. *arXiv*, 2021.

Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun et Hung-Kuo Chu : DuLa-Net : A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3358–3367. IEEE Computer Society, 2019. ISBN 978-1-72813-293-8.

Wenyan Yang, Yanlin Qian, Francesco Cricri, Lixin Fan et Joni-Kristian Kamarainen : Object Detection in Equirectangular Panorama. *In 2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2190–2195. IEEE Computer Society, 2018. ISBN 978-1-5386-3788-3.

Boram Yoon, Hyung-il Kim, Gun A. Lee, Mark Billinghurst et Woontack Woo : The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. *In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 547–556. IEEE, 2019. ISBN 978-1-72811-377-7.

Jacob Young, Tobias Langlotz, Matthew Cook, Steven Mills et Holger Regenbrecht : Immersive Telepresence and Remote Collaboration using Mobile and Wearable Devices. *IEEE Transactions on Visualization and Computer Graphics*, 25 (5):1908–1918, 2019. ISSN 1941-0506.

Jacob Young, Tobias Langlotz, Steven Mills et Holger Regenbrecht : Mobileportation : Nomadic Telepresence for Mobile Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):65 :1–65 :16, 2020.

Karim Youssef, Sherif Said, Samer Al Kork et Taha Beyrouthy : Telepresence in the Recent Literature with a Focus on Robotic Platforms, Applications and Challenges. *Robotics*, 12(44):111, 2023. ISSN 2218-6581.

Kevin Yu, Gleb Gorbachev, Ulrich Eck, Frieder Pankratz, Nassir Navab et Daniel Roth : Avatars for Teleconsultation : Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, 2021. ISSN 1941-0506.

Matteo Zanetti, Alessandro Luchetti, Sharad Maheshwari, Denis Kalkofen, Manuel Labrador Ortega et Mariolino De Cecco : Object Pose Detection to Enable 3D Interaction from 2D Equirectangular Images in Mixed Reality Educational Settings. *Applied Sciences*, 12(1111):5309, 2022. ISSN 2076-3417.

Wei Zeng, Sezer Karaoglu et Theo Gevers : Joint 3D Layout and Depth Prediction from a Single Indoor Panorama Image. *In Computer Vision - ECCV 2020*, volume 12361 de *Lecture Notes in Computer Science*, pages 666–682. Springer International Publishing, 2020. ISBN 978-3-030-58516-7.

Emin Zerman, Cagri Ozcinar, Pan Gao et Aljosa Smolic : Textured Mesh vs Coloured Point Cloud : A Subjective Study for Volumetric Video Compression. *In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

Dengsheng Zhang et Guojun Lu : Study and evaluation of different Fourier methods for image retrieval. *Image and Vision Computing*, 23(1):33–49, 2005. ISSN 0262-8856.

Junzhe Zhang, Daxuan Ren, Zhongang Cai, Chai Kiat Yeo, Bo Dai et Chen Change Loy : Monocular 3D Object Reconstruction with GAN Inversion. *arXiv*, 2022a.

Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong et Baining Guo : VirtualCube : An Immersive 3D Video Communication System. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2146–2156, 2022b. ISSN 1941-0506.

Yajie Zhao, Qingguo Xu, Weikai Chen, Chao Du, Jun Xing, Xinyu Huang et Ruigang Yang : Mask-off : Synthesizing Face Images in the Presence of Head-mounted Displays. *In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 267–276, 2019.

Chao Zhou, Zhenhua Li et Yao Liu : A Measurement Study of Oculus 360 Degree Video Streaming. *In Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 27–37. Association for Computing Machinery, 2017. ISBN 978-1-4503-5002-0.

Fuqiang Zhou, Bin Peng, Yi Cui, Yexin Wang et Haishu Tan : A novel laser vision sensor for omnidirectional 3D measurement. *Optics & Laser Technology*, 45:1–12, 2013. ISSN 0030-3992.

Katja Zibrek, Sean Martin et Rachel McDonnell : Is Photorealism Important for Perception of Expressive Virtual Humans in Virtual Reality ? *ACM Transactions on Applied Perception*, 16(3):14 :1–14 :19, 2019. ISSN 1544-3558.

Katja Zibrek et Rachel McDonnell : Social presence and place illusion are affected by photorealism in embodied VR. *In Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '19, pages 1–7. Association for Computing Machinery, 2019. ISBN 978-1-4503-6994-7.

Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer et Tony Tung : Data-Driven 3D Reconstruction of Dressed Humans From Sparse Views. *In 2021 International Conference on 3D Vision (3DV)*, pages 494–504, 2021.

Nikolaos Zioulis, Dimitrios Alexiadis, Alexandros Doumanoglou, Georgios Louizis, Konstantinos Apostolakis, Dimitrios Zarpalas et Petros Daras : 3D tele-immersion platform for interactive immersive experiences between remote users. *In 2016 IEEE International Conference on Image Processing (ICIP)*, pages 365–369, 2016.

Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas et Petros Daras : OmniDepth : Dense Depth Estimation for Indoors Spherical Panoramas. *In Computer Vision - ECCV 2018*, Lecture Notes in Computer Science, pages 453–471. Springer International Publishing, 2018. ISBN 978-3-030-01231-1.

Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Niessner, Reinhard Klein et Andreas Kolb : State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum (EG STAR)*, 37(2):625–652, 2018.