WISSENSCHAFTLICHE ARBEITEN DER FACHRICHTUNG
GEODÄSIE UND GEOINFORMATIK DER LEIBNIZ UNIVERSITÄT HANNOVER
ISSN ????-????

Nr. ???

Control of Walking behaviour in Shared Spaces using Augmented Reality

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation von
Dipl.-Ing. Vinu Kamalasanan
geboren am 05.02.1989 in DUBAI,UAE

HANNOVER 2024

Diese Arbeit ist auch veröffentlicht in:
DEUTSCHE GEODÄTISCHE KOMMISSION bei der Bayerischen Akademie der Wissenschaften
Reihe C, Dissertationen, Heft Nr. ???, München 20??
ISBN ???, ISSN ???
www.dgk.badw.de

Prüfungskommission:

Vorsitzender:	Prof. Dr.-Ing Claus Brenner
Referent:	Prof. Dr.-Ing. Monika Sester
Korreferenten:	Prof. Dr.-Ing. Jörg.P.Müller
	Prof. Dr.-Ing. Ingo Neumann
Tag der Promotion:	06.08.2024

Abstract

Walking is by far the most important form of active transportation that helps to both explore the environment and also socialise with others while navigating a street. However, with the increasing number of vehicles, the safety distance that pedestrians are expected to maintain with other fellow pedestrians, cyclists and cars has reduced. This has raised safety concerns, especially in mixed-traffic urban designs like Shared Spaces. Shared traffic spaces are mixed environments where the physical separation between walking pedestrians and other road participants (like cyclists and vehicles) is reduced. The mix of different road users and fewer rules is expected to increase communication and interaction between pedestrians and others. The streets designed using the sharing principle are characterised by the removal of traffic signals and cyclist lanes while adding street furniture to promote pedestrians space utilisation. While such designs are mainly focused on reducing the dominance of vehicles, pedestrians especially the elderly and disabled, feel less safe using such spaces. This is because they find it hard to estimate threats and potential collisions when crossing paths with fellow pedestrians, cyclists or vehicles in the scene. Moreover, when considering pedestrians outside the elderly population, little research has addressed how AR could be used to improve the perception of safety and enhance walkability in public spaces.

To enhance pedestrian safety, this research aims to leverage Augmented Reality (AR) to influence pedestrian path choices and walking behavior during collision avoidance. To achieve this, the thesis first reviews related work on the use of visual information to manipulate walking paths. Then a scene perception pipeline is proposed, implemented and evaluated. Once it has been proven that it is possible to use the AR headset for scene perception - to detect and track the movements in the walker's surroundings, the visualization of the future path of the neighboring pedestrians is used to influence the AR headset user's choice of route.

To demonstrate a proof of concept in applying motion perception to AR walking influence, this thesis uses the Hololens and its RGBD sensors. A workflow is designed and implemented consisting of detection and tracking to monitor the movements of individuals in the environment who may potentially come into conflict. For the pipeline, 3D pedestrian detection algorithm is used to localise surrounding pedestrians while pedestrian tracking is implemented to associate the detection's and approximate their walked trajectories in the scene. Then a concept to visualise the future path in order to influence the ego user is presented. For this, the design and implementation of candidate visualisations and a subsequent user study was completed to study preferred walking paths to destinations while seeing the future motion of others augmented with AR. The preferred paths were then compared from a safety perspective to evaluate the implications of seeing future path visualisation. The findings indicated that people prefer to walk longer trajectories and safer paths around conflict points with future information. Also to study other AR influences that could affect walking, two methods of influences - AR for virtual traffic infrastructure implementation and AR to represent moving and crossing traffic agents are evaluated. Both studies when evaluated using surrogate safety measures pointed towards AR content playing a role in promoting safer walks.

The results of this thesis show that the use of both an AR device and appropriate visualisation can successfully detect and track nearby pedestrians and visualising future path influences the collision avoidance behaviour. While the results from the study support and

enhance safer walking, the findings are a foundation to understand how pedestrians with AR interact and collaborate with cyclists and autonomous vehicles.

Keywords: Shared spaces, motion conflicts, pedestrian detection and tracking in AR, future path

Kurzfassung

Zu Fuß zu gehen ist bei weitem die wichtigste Form der aktiven Fortbewegung, die sowohl dabei hilft, die Umgebung zu erkunden als auch beim Befahren der Straße mit anderen in Kontakt zu kommen. Mit der zunehmenden Zahl von Fahrzeugen hat sich jedoch der Sicherheitsabstand verringert, den Fußgänger zu anderen Fußgängern, Radfahrern und Autos einhalten müssen. Dies hat Sicherheitsbedenken aufgeworfen, insbesondere bei gemischten Verkehrsgestaltungen wie Shared Spaces. Shared-Verkehrsräume sind gemischte Umgebungen, in denen die physische Trennung zwischen Fußgängern und anderen Verkehrsteilnehmern (wie Radfahrern und Fahrzeugen) verringert ist. Die Mischung verschiedener Verkehrsteilnehmer und weniger Regeln sollen die Kommunikation und Interaktion zwischen Fußgängern und anderen erhöhen. Die nach dem Sharing-Prinzip gestalteten Straßen zeichnen sich durch die Entfernung von Ampeln und Radfahrstreifen aus, während Straßenmöbel hinzugefügt werden, um die Nutzung des Raums durch Fußgänger zu fördern. Während solche Gestaltungen hauptsächlich darauf ausgerichtet sind, die Dominanz von Fahrzeugen zu verringern, fühlen sich Fußgänger, insbesondere ältere und behinderte Menschen, bei der Nutzung solcher Räume weniger sicher. Dies liegt daran, dass es ihnen schwerfällt, Gefahren und mögliche Kollisionen einzuschätzen, wenn sie den Weg anderer Fußgänger, Radfahrer oder Fahrzeuge kreuzen. Auch wenn man Fußgänger außer der älteren Bevölkerung betrachtet, gibt es wenig Arbeit darüber, wie AR genutzt werden könnte, um das Sicherheitsgefühl zu verbessern und die Begehbarkeit öffentlicher Räume zu verbessern.

Um Fußgänger aus Sicherheitsgründen zu unterstützen, beabsichtigt diese Forschung, Augmented Reality (AR) zu nutzen, um die Wahl des Fußgängerwegs und das Gehverhalten bei der Kollisionsvermeidung zu beeinflussen. Zu diesem Zweck überprüft die Arbeit zunächst verwandte Arbeiten darüber, wie visuelle Informationen zur Manipulation von Gehwegen verwendet wurden, und schlägt eine Szenenwahrnehmungspipeline als wesentliche Komponente vor, um einen Fußgänger in einer Szene zu beeinflussen. Sobald nachgewiesen wurde, dass es möglich ist, das AR-Headset zur Szenenwahrnehmung zu verwenden – um die Bewegungen in der Umgebung des Fußgängers zu erkennen und zu verfolgen –, könnte die Visualisierung des zukünftigen Weges der benachbarten Fußgänger verwendet werden, um die Routenwahl des AR-Headset-Benutzers zu beeinflussen.

Um einen Proof of Concept bei der Anwendung der Bewegungswahrnehmung auf den Einfluss von AR-Gehen zu demonstrieren, verwendet diese Arbeit HoloLens und seine RGBD-Sensoren, um einen Workflow zu entwerfen und zu implementieren, der aus Erkennung und Verfolgung besteht, um die Bewegungen der Personen in der Umgebung zu erfassen, die potenziell in Konflikte geraten könnten. Für die Pipeline wird die 3D-Fußgängererkennung verwendet, um umgebende Fußgänger zu lokalisieren, während die Verfolgung implementiert wird, um die Erkennungen zu verknüpfen und ihre Gehbahnen in der Szene zu approximieren. Anschließend wird ein Konzept zur Visualisierung des zukünftigen Pfads zur Beeinflussung des Ego-Benutzers vorgestellt. Hierzu wurde ein Entwurf und eine Implementierung möglicher Visualisierungen und eine anschließende Benutzerstudie durchgeführt, bei der bevorzugte Gehwege zu Zielen während des Sehens einer Szene mit zukünftigen Bewegungen anderer mit AR erweitert wurden. Die bevorzugten Pfade wurden dann aus Sicherheitsgründen verglichen, um die Auswirkungen der Visualisierung zukünftiger Pfade zu bewerten. Die Ergebnisse zeigten, dass Menschen es vorziehen, längere und sicherere Wege um Konfliktpunkte herum zu gehen, mit zukünftigen Informationen. Um auch andere AR-Einflüsse zu untersuchen, die das Gehen beeinflussen könnten, werden zwei Einflussmethoden untersucht – AR für die Implementierung virtueller Verkehrsinfrastruktur und AR zur Darstellung von

sich bewegenden und kreuzenden Verkehrsteilnehmern. Beide Studien, bei denen sie anhand von Ersatzsicherheitsmaßnahmen ausgewertet wurden, zeigten, dass AR-Inhalte eine Rolle bei der Förderung sichererer Spaziergänge spielen.

Die Ergebnisse dieser Arbeit zeigen, dass der Einsatz eines AR-Geräts und geeigneter Visualisierungen erfolgreich dazu genutzt werden kann, Änderungen der bevorzugten Gehwege und der Art und Weise herbeizuführen, wie eine Person mit einer Kollision konfrontiert wird. Während die Ergebnisse der Studie dazu verwendet werden könnten, sichereres Gehen zu unterstützen und zu verbessern, könnten die Erkenntnisse auch als Grundlage verwendet werden, um zu verstehen, wie Fußgänger mit AR mit Radfahrern und autonomen Fahrzeugen interagieren und zusammenarbeiten würden.

Schlagworte: Gemeinsam genutzte Räume, Bewegungskonflikte, Fußgängererkennung und -verfolgung in AR, zukünftiger Weg

Contents

1 Introduction

Walking has been an important form of active transportation mode since early civilisations. Along with cycling, this active mode is found to differ with cultures from 6% in North America (USA) to 46% in Europe (Netherlands). Navigating on the foot not only improves one's health but also enhances social presence and initiates verbal interactions. When people see other familiar and friendly people in the neighbourhood, they are more interested in walking and might take longer navigation paths and routes (Gallagher et al. 2010). Walking in outdoor spaces, amongst other factors, would highly depend on the perceived level of safety and increased feeling of confidence and priority. Aesthetics and greenery of outdoor spaces would act as enablers and motivate walking; while the noise of moving vehicles could be a deterrent (Andersson et al. 2023). Furthermore, clean and well-maintained streets with appealing sights support the apprehension of space as a better walkable landscape (Lees et al. 2007, Grant et al. 2010).

1.1 Motivation

With the advent of motor vehicles and urbanisation, the importance of walking and the priority of pedestrians in traffic has reduced drastically. The increasing number of cars has lowered their safety distances to pedestrians when sharing their walking spaces. This has made unsafe and risky pedestrian traffic encounters more frequent. Also with traffic signals now in place, there are fewer opportunities for vulnerable road users like pedestrians and cyclists to communicate with each other (via gesture and gaze) to negotiate priority. This further removes the purpose of walking as a way to explore outdoor spaces. As a result of this, people tend to use public spaces only when necessary, as they perceive it slightly dangerous otherwise. Furthermore, they might tend to walk shorter paths due to the fear of traffic encounters. Such fears might discourage elderly persons from using walkable spaces due to their inability to estimate threats with reduced cognition and increasing ageing (Distefano et al. 2021).

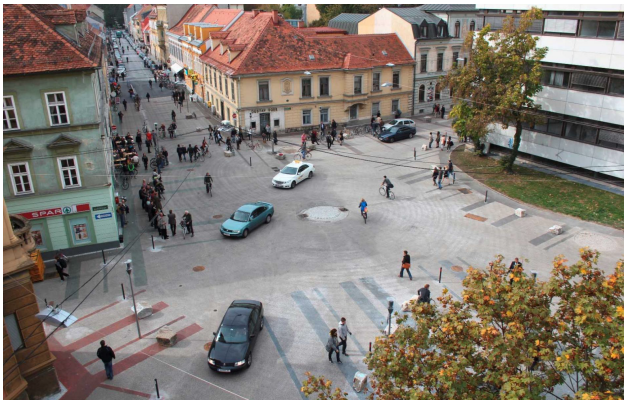


Figure 1.1: A shared space in Sonnenfelsplatz, Graz, Austria. ©Helke Falk

Safety reports have pointed towards the relative decrease in the number of people walking (Do 2002) over the years. In one such report (Boarnet et al. 2005) that analysed the commuting behaviour of school children, the recorded drop in walking rate was as high as 67% over three decades in the US. These figures were closely mapped to the risks of school travel. Children opted to commute with private cars due to increased fear of pedestrian and cyclist injuries. Similar trends that reduced walking could be observed when traffic influences and built environment were accounted for in *neighbourhood walkability* studies. Higher walking rates were observed in adults (Van Dyck et al. 2009) when safer neighbourhoods promoted people to walk more than the less safe streets. As environmental factors have played an important role in increasing the feeling of pedestrian safety, traffic planners have taken this into account when designing streets to improve walking. This has resulted in newer street designs being proposed and improved from time to time.

The concept of livable streets for pedestrians was introduced by Appleyard (Appleyard 1980) as an alternative to car-centric street designs. The concept was focused on redefining cities that are often viewed as dangerous, polluted and noisy by residents and visitors. Such designs and similar street layout principles like the *shared spaces* (Figure 1.1) aim at reducing vehicle speeds and improving pedestrian priority in traffic spaces. Shared Spaces are designed to enable pedestrians to move freely while reducing the physical segregation between them and other road users sharing the same movement space. Such road designs built to support pedestrian priority, have been reported to reduce the number of collisions involving pedestrians, cars and moped drivers (Alink 1990).

By minimizing traffic control, shared spaces integrate the different traffic participants forcing traffic decisions to be based on road-user interactions (hand signals and gestures) and mutual respect for each other's priority (Hamilton-Baillie 2008). Such urban designs are characterised by the removal of kerbs and footways, mixing the different road users. Then people might pay more attention when crossing each other's paths. Even physical retail marketplaces like a farmer's market could be categorised as a shared space. Such an environment would contain street furniture that is in place for customers while also having vehicles and cyclists moving nearby. This then creates situations where moving persons and crowds come in close contact to passing cyclist or cars which might force them to slow down. Hence such design concepts could promote pedestrian walks while reducing the dominance of cars with drivers forced to yield to crossing pedestrians.

1.2 Current Issues

Even when such mixed traffic designs have contributed to improving social interactions and the sense of safety for pedestrians (Sauter and Huettenmoser 2008), there has been widespread criticism against their acceptance. As everyone has to pay attention while crossing, the elderly and disabled feel more stressed in using such spaces. Also, the absence of clear separation between pedestrian, cyclist and vehicle paths prompt the elderly to believe that they can no longer use the space independently (Thomas 2008). To mitigate these safety issues and to support pedestrian mobility, the Ramboll Nyvig report (Nyvig 2007) proposed alternative designs to enhance shared spaces. The report advocated the creation of '*safe spaces*' within shared spaces where vulnerable pedestrians would remain away from cars. The proposed 'safe space' recommended adding separate footways catering to only the elderly with the remaining space available to be shared by motorists, cyclists and other pedestrians. Another report (Lawson et al. 2022), emphasised having better pedestrian-

friendly safety regulations in place for pedestrians. Both these proposals contradict the true purpose of shared spaces, where the mix of different participants and the resulting traffic negotiations are expected to play a significant role.

While safety concerns of the elderly have been a topic of debate, before-after shared space analysis, (Kaparias et al. (2015) and Fu et al. (2019)) have further pointed out current constraints that still exist to walking and the inherent pedestrian priority issues that need to be resolved. Such research reports based on movement analysis give a better insight into those traffic agents who have benefited more from the creation of mixed street design. For example in the analysis made for a temporary shared space setting in (Batista and Friedrich 2022a), the study revealed that of all the road users, cyclists and cars repeatedly used only the same travel path in the mixed traffic space. Pedestrians, on the other hand, used the complete navigation space and showed a tendency to walk longer paths including a few detours. This observation pertaining to pedestrian space utilisation in the study was highly dependent on vehicular density and design of the layout (Batista and Friedrich 2022b). For instance, people navigated the entire shared space only when motor vehicles were lesser in number. As the vehicular density increased, walking behaviours amongst pedestrians differed with people preferring to use only specific navigation paths. Also when the layout of the street was more narrow, pedestrians again used the complete space; which changed with walking concentrated to safer pedestrian zones in case of higher volumes of vehicles on the street. Another interesting finding in the study was the disparity of yielding behaviour in interactions involving pedestrians and vehicles (especially the right of way behaviour). In three of the five shared spaces, pedestrians mostly gave the right of way to vehicles in the event of a priority conflicts. Hence traffic priority and the resulting space utilisation of movement spaces were highly dependent on those interaction situations that brought pedestrians in direct contact to crossing vehicles. While not much work has been done in literature to enhance space utilisation, few concept ideas and proposals to influence movement and enhance traffic priority has been proposed recently.

Movement influences in traffic spaces - eHMI and future concepts To contribute towards pedestrian safety and the resulting priority while interacting with vehicles, external Human Machine Interface (eHMI) concepts have been tested in traffic studies. Such interfaces communicate crossing decisions when interacting with an autonomous car (e.g. conveying their priority to cross first). For this, ideas have suggested that the vehicles should either show / project messages whenever there is a priority confusion to support safer pedestrian-vehicle interactions. This could include projecting green or red lights or by showing text messages of who could cross, as shown in Figure 1.2. Most of these communicative eHMI designs that have been researched in both academia (Busch et al. 2018, Dey et al. 2020, Rouchitsas and Alm 2019) and industry (Bazilinskyy et al. 2019) have mainly focused on traffic priority negotiations via signalling.

In most designs, the communications differ based on the placement of the message source and how the information is shared with the eHMI. The message could be conveyed using projections onto the road surface, by emitting light via an embedded display on the vehicles and by communicating priority to a nearby traffic signal or pedestrian smartphone (Colley et al. 2017a, Colley et al. 2022, Holländer et al. 2020, Ackermann et al. 2019, Busch et al. 2018). While most of these approaches have focused on communicating what the car would like to do when it is confronted with a pedestrian during a crossing, such interfaces assume that the pedestrian would comply and follow the recommendation given by the car. This

Figure 1.2: Examples of several eHMI concepts how an AV would communicate the intention to other pedestrians around (Photo : Dey et al. 2020).

would mean that if the AV intends to give the right of way, then the pedestrian crosses, but otherwise not. This also means that vehicles would still remain the dominant road user in traffic spaces, with pedestrians only communicating their willingness (Dey et al. 2021) to cross.

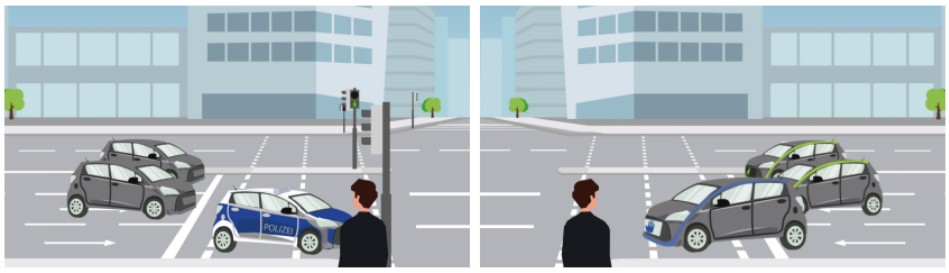


Figure 1.3: Example intersection with multiple vehicles signalling in green allowing the person to cross the intersection (Löcken et al. 2023).

Another interesting aspect of the proposed eHMI-mediated traffic negotiation is its ability to achieve collaborative actions with other agents. In Löcken et al. (2023), a eHMI based virtual traffic signalling alternative was suggested when self-driving cars communicated with other AVs nearby. This resulted in collective behaviours amongst different AVs as they got closer to an intersection. For instance, when the vehicles approached the junction (Figure 1.3), all the vehicles would stop at once and project a green light in unison that would then allow fellow pedestrians to cross. The nature of such one-to-one or one-to-many AV dominated interactions however does not address the issue of pedestrian empowerment or eHMIs mediating as a technology to enforce pedestrian priority at a traffic interaction.

Our earlier works towards increasing the priority of pedestrians in traffic spaces (Li et al. 2022b) focuses on a vision of using Augmented Reality (AR) to enhance pedestrian priority via collective actions like grouping (Figure 1.4). The idea in the work emphasised that when virtual paths were shown individually on AR glasses, people decided to walk and cross differently than otherwise. The concept exploits the use of virtual lanes to invite pedestrians to form groups. Seeing a common virtual walkable crossing path using an AR headset or street projection could then prompt people to move and interact differently with cars, cyclist or even other pedestrains. For this, people walking nearby potential group members have to be tracked using intelligent sensors and then suggested to walk together

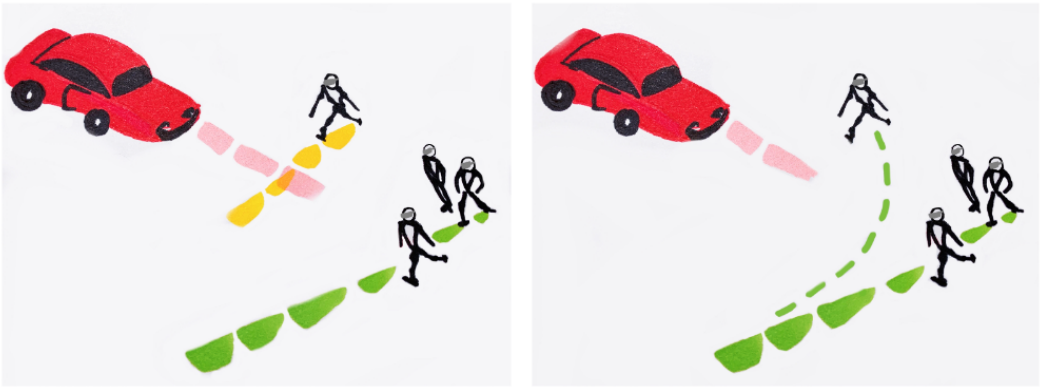


Figure 1.4: (left) A single pedestrian wants to cross the shared road while being confronted with an AV. (right) The appearance of a virtual lane prompts the AV to stop, as pedestrians get a higher priority for their willingness to join in a group and walk together.

as a group. But for such an idea to be realised, two key requirements have to be met. Firstly, it should be proven that its possible to detect and track nearby walking persons using a AR headset or any medium that supports augmentation and secondly, any virtual recommendations suggested based on surrounding motion should be proven to influence walking behaviour.

However to realise the first prerequisite, a motion perception algorithm would be required to detect and track surrounding pedestrian motion in the scene. As the idea would be to use an AR overlay in the subsequent step, the perception pipeline should be demonstrated using an AR headset and its embedded sensors. Furthermore while proving the feasibility of the AR device for motion perception is important, equally needed is the requirement to benchmark how accurately such a device can infer surrounding walking movement. This is because the AR overlays that would be shown based on pipeline outputs depend on the estimated positions and accuracy of the motion perception inputs. Mobile AR (Höllerer and Feiner 2004) visual overlays have to be accurate to a high degree so that the graphics rendered looks realistic and registered to the real world in natural light settings.

For the second prerequisite, AR content that will be shown based on inferred motion should be able to influence walking path choices of the content viewer. For example, if the AR medium initially shows the future walking steps of nearby persons and then communicates overlapping virtual paths; the resulting path choice influences to the viewer from such a visual approach has to be studied. This will help to better understand what visual information will prompt different walking styles. For instance, whether people are choosing to walk longer paths or shorter trips when they already see the future motion of persons moving in front of them. Also, as the output from the motion perception algorithm would be the input to the AR visual overlay; the visual information that is communicated should be different for the the different perception inferences. This should influence walking differently when showing different visual representations. This would mean that if the pipeline predicts future motion *confidently* or with a level of *uncertainty*, this should be visually communicated with AR. Different visual representation of the same future information might prompt different walking behaviours that needs to be then evaluated. Also from a visual interpretation perspective, while showing a future path could be one approach, other forms of visual influences (e.g., virtual avatars or virtual traffic signals) that could influence different walking behaviors should be further explored.

1.3 Goal of the Thesis and Scientific Contribution

The goal of this thesis is to explore from a safety perspective, the impact of an AR device and their visualization in influencing human path choices. Hence as the primary focus, we try to observe how persons would avoid a collision when seeing AR signals that would either assist, instruct or manipulate them to make different navigation decisions. For this, the thesis focuses on the following research questions :

- Can an AR device and its sensor be used to understand (walking) motion surrounding the headset user ?
- Can movement in the immediate vicinity of the headset be detected and tracked with a high level of accuracy ?
- Could visualising the future information of the nearby traffic agent movement prompt safer collision avoidance behaviours ?
- Does the visual representation of this future information prompt shorter or longer walking paths during a collision avoidance situation ?

To answer these questions, we primary focus on the AR headset Hololens, motion perception using its sensor and how a Hololens user could be influenced on his/her navigation path with 3D visualisations based on the surrounding scene movement. The Hololens user in our work would then be synonymous to a traffic pedestrian walking in a shared space. While this thesis intends to demonstrate a proof of concept on how AR could be instrumental in influencing walking, it does not prove walking manipulation for outdoor spaces. The realisation of AR influences for pedestrians in shared spaces is however currently limited by the acceptability and technical limitations of AR technology. This will not be covered under the scope of this current work.

In addition to answering the above research questions, the contribution of this thesis are :

- A first proof of concept study to explore the influence of AR stop-and-go signals as virtual traffic infrastructure to influencing walking.
- The work presents a prototype framework that is used to study manipulation of walking resulting from visualising moving 3D traffic agents (cyclists) as avatars in mixed reality.
- Trajectory based analysis with surrogate safety measures have been applied for the first time to study walking influences.

1.4 Structure of Thesis

This thesis is structured as follows:

Chapter 2 will provide a background on Shared Spaces, Visualisations, Motion perception using object detection and tracking. Lastly, User study design principles and how methods are chosen when conducting a study are also explained.

Chapter 3 discusses the related works on active, passive and external influences that influence walking behaviour. The research gaps are further identified in this section.

Chapter 4 discuss on the datasets and evaluation metrics used in all the experiments in this thesis.

Chapter 5 demonstrates how scene perception is achieved using the Hololens sensors by focusing on the implementation of both pedestrian detection and tracking using the RGBD sensors.

Chapter 6 summarises the different experiments that were completed on the Hololens data to test its capability in detecting and tracking pedestrians accurately. The section also discusses the current limitations of the motion perception pipeline that is realised.

Chapter 7 explains the method and inferences made on visualising future path of others to influence path choices of a pedestrian seeing it and navigating a conflict.

Chapter 8 summarises the findings of how AR could influence walking based on *other* sources like a virtual traffic infrastructure or moving avatars that would either instruct or manipulate the crossing path of a walking person.

Chapter 9 discusses the results from the different methods proposed in the research. Along with evaluation of the motion pipeline and the future visualisation, future works that could be addressed within the scope of this thesis are also identified.

2 Background

This chapter presents some basic concepts and explanations that are necessary for understanding the dissertation work. As stated in Chapter 1, this thesis focuses on applying AR visualisation techniques to influence the walking behaviours of pedestrians in shared spaces. Thus in this chapter, first a detailed description of shared spaces designs and pedestrian interactions is introduced. Then some fundamentals on visualisation and 3D representation are detailed. To present components of a perception pipeline, the Hololens and 3D sensing is explained. Object detection and tracking approaches that are relevant to this thesis are then further elaborated. Lastly, the steps to designing a user study to investigate effective visualisations techniques are covered in the final section.

2.1 Shared Spaces

"Traffic is a sort of secret window onto the inner heart of a place, a form of cultural expression as vital as language, dress or music"- (Vanderbilt 2009, p. 216)

Traffic and urban spaces are vital to our society not just to facilitate movement from one place to another, but also to encourage social encounters amongst pedestrians while ensuring accessibility to local businesses. However life on the streets is highly related to the characteristics of its built environment. While traffic signs and signals regulate priority between cars, cyclist and pedestrians; environmental features (e.g., layout, cycle lanes and walking paths) segregate different road users. These traffic features are also to a larger extend responsible for influencing and deciding on how and when pedestrians would cross (or interact) while walking and navigating with vehicles and cyclists.

(a) Friesland, the Netherlands: A five way intersection with a shared level, traffic lights, sign and markings removed (Photo: Hamilton-Baillie 2004).

(b) Rijksstraatweg, the Netherlands: Shared level surface with paving creates a town square with seating and restaurants (Photo: Sutcliffe 2009).

Figure 2.1: Shared spaces illustrations from the Netherlands.

Shared spaces is a design principle introduced by the Dutch traffic engineer Hans Monderman (Space 2005); focused to improve the quality of conventional streets by enhancing priority of the vulnerable road users (pedestrians and cyclists). This has been targeted to be achieved

by reducing the vehicular dominance of traffic spaces. In comparison to current street designs, the concept attempts to minimize traffic control and segregation between pedestrian and other road users. This is done by removing features such as kerbs, road surface marking, traffic signs and traffic lights. With such measures, the idea is to improve both the traffic interaction and communication between different road users using the streets. Since its introduction, the shared space has received considerable attention worldwide including street designs both in the USA and New Zealand. The Figure 2.1 shows the two shared space intersections that are currently being used as mixed traffic spaces.

2.1.1 Walking and Shared Spaces

The *fluidity of movement* and *sensory experiences* are important aspects associated with the mobility mode walking (Delaney 2016). The experience of a "fluid motion" is linked to the motion continuity of the walking paths and the accompanying satisfaction in maintaining it. This somehow is mapped to comfort as manoeuvring their body is part of the pedestrian walking experience. As a person would walk in a shared traffic space, there could be confrontations with other pedestrians groups, cyclist and vehicles who would cross their walking paths. There could also be situations when the person is forced to stop and give up priority due to reckless cyclist and other persons with unfriendly behaviours. Unexpected stops could then hamper motion continuity, inducing frustration and unease while ending the fluid journey that was in progress during walking.

Walking by foot being the slowest mode of active transportation has its own "sensory" benefits. Due to its relaxed pace, walking captures 'treasured views' as visual senses predominate in this mode of transport. Furthermore due to the steady speed of things moving past, pedestrians tend to be more sensitive to the surrounding and also on their choices on whom to share their walking space with. While sharing walking path with strangers could happen at random in public spaces, the movement rhythm created with others in the process could initiate a feeling of closeness and bonding (Coleman and Collins 2020). This could result in informal eye-contacts which might mostly be confrontational and less companionable. Traffic relationships in cities as stated by Hannerz (Hannerz 1980) 'are a pure form of meetings among strangers, a result of the crowding of large numbers of people in a limited space'.

2.1.2 Pedestrian Interactions in Shared Spaces

While walking in shared spaces, pedestrians might cross paths with other road users more frequently. Every crossing maneuver would then involve either of the involved participants either adjusting their path or speed of motion. The outcome of every pedestrian interaction observed in such cases would depend on how the involved traffic participants avoided a space-sharing conflict (Markkula et al. 2020). Conflict avoidance amongst many factors depend on the type of traffic participant (pedestrian, cyclist or vehicle) involved in the interaction and how they confronted the pedestrian in the scene (Lehsing and Feldstein 2018). Markkula et al. (2020) states that there are limited number of ways that two road users can approach a conflict space with interactions generalized to five prototypical space sharing scenarios - obstructed paths, merging paths, crossing paths, unconstrained and constrained head-on paths (Figure 2.2). Also when more than two road-users would be involved, the resulting interaction could results from multiple prototypes of the above stated scenarios being applied

simultaneously. In the rest of this section, we primarily focus on pedestrian-pedestrian and pedestrian-cyclist interactions that has mainly been the focus of this thesis.

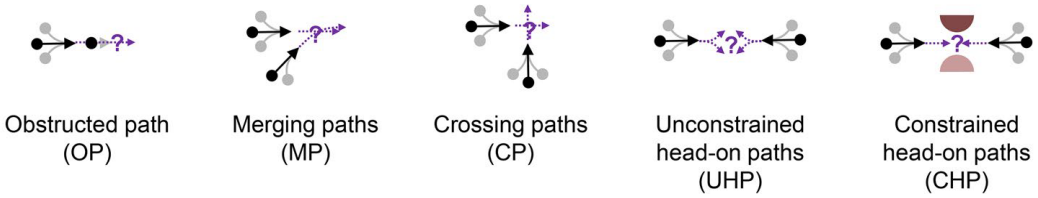


Figure 2.2: Different types of conflicts borrowed from (Markkula et al. 2020), illustrate the potential interactions that could arise between a pedestrian and other road user (autonomous car). Each arrow represents either a pedestrian or vehicle conflicting the other.

Pedestrian - Pedestrian Interactions: Pedestrians avoid collisions with other pedestrians either by longitudinal and lateral evasive maneuvers (Meerhoff et al. 2018) when crossing paths. Based on how users would meet in a shared space and how they negotiate priority amongst themselves, three configurations/situations were noted in shared spaces (Jensen 2010). The first *frontal meeting* observed in the work, is a situation common in two way traffic where pedestrians could meet others face to face. This interaction configuration is similar to the unconstrained head-on paths (UHP) and constrained head-on paths (CHP) illustrated in Figure 2.2. The second "orthogonal meeting" happens when the surrounding street geometry (T shaped intersection e.t.c.,) might prompt orthogonal meeting between pedestrians and other road users. The merging paths (MP) and crossing paths (CP) in Figure 2.2 could be considered orthogonal meeting cases. The third meeting configuration "parallel meeting " might mostly result while taking parallel paths. A situation that might arise when cyclists overtake a walking pedestrians from behind (OP from Figure 2.2).

While conflicts would result when pedestrians meet based on any of the three cases discussed above, either of the pedestrians would negotiate to avoid a future collision. Some of the particularly interesting negotiations that could be observed during pedestrian-pedestrian interactions are mentioned in Table 2.1.

Type	Interaction
Confusion	No one gives a clear signal over yielding their right of way
Both giving in	Both person give in a little and pass each other by
Zigzag turning	The pedestrian performs a zigzag walking motion past the other person
Stop to pass	One of the pedestrians would stop so that the other passes by
Group passing a pedestrian	The members come closer making space for person to pass
Group letting a stranger in	The group splits with the pedestrian passing between them

Table 2.1: Pedestrian-pedestrian negotiations in shared spaces (Jensen 2010).

Pedestrian - Cyclists Interactions: Due to the relatively faster speed of movement, pedestrians interactions with cyclists could be more uncertain; characterised by trajectories with both speed or path deviations over a shorter interaction time. This makes studying pedestrian-cyclist interactions different from pedestrian-pedestrian movements and how their motion trajectories differ. Even when recent shared spaces datasets (Mukbil et al.

2023) as in Figure 2.3 have captured interactions between cyclist and pedestrians, lesser works have tried to address the different movement configurations (as illustrated in Figure 2.2) that would result in a conflict between pedestrian and space sharing cyclist.

Figure 2.3: Cyclist interactions with other pedestrians and vehicles as illustrated by CTV-Dataset (Mukbil et al. 2023).

So far, most studies have focused on observing and analysing the interaction between pedestrians and cyclists. A recent study investigated the interaction behaviour of a group of pedestrians with a cyclist on a shared road (Huang et al. 2021) via a series of controlled experiments. However, the work only focused on the macroscopic influence of a pedestrian crowd on individual cyclist's paths. Further, the generalizability of the study results was limited by the particular study setting and group behavior. Another controlled experimental study investigated the interaction of cyclists (Yuan et al. 2018) and restricted the research to bicycles only. The work did not include pedestrians as interaction partners and their potential interaction influences.

Hence studying pedestrian interactions with both other persons and cyclists have been completed using either real world motion data or by conducting controlled experiments. As pedestrians would move differently based on how and with whom (other cyclist or person) interactions took place, careful considerations should be made when designing data capture experiments. This includes emphasis on those scenarios on how persons would walk when they share their walking space and cross paths with others.

2.2 Visualisations

From weather maps to graphics applied to transportation systems and in-flight entertainment; visualisation is part of our everyday life. Visualisation engages the sensory apparatus-*vision*, as well as the processing power of the human mind in interpreting and understanding information. It is a simple and effective medium of communication with which even complex and high-dimensional data (e.g., volumetric information) can be easily represented and understood. *"Visualisation is the act or process of interpreting in visual terms or of putting into visual form"*- (Merriam-Webster 1990).

Visualisation is necessary to make sense of the large amount of information in the present day world. Without the use of proper visualisations, most of the hidden phenomena that are present in our immediate surroundings could go unnoticed and might be overlooked. Even in some situations like daily driving, due to the overload of tasks; we often tend to miss out on details (e.g., a person crossing from the edge of the street) which might later turn out catastrophic. A proper visualisation in such safety critical situations can be very helpful. Representing visual clues shown in 3D can helps to paint a better mental picture of the threatening event, which along with attracting user attention would help in averting accidents.

Due to its focus on representing information, visualisation is often confused with computer graphics and imaging. As per the distinctions presented in (Schroeder et al. 1998), visualisation differentiates itself from graphics in three ways:

- The dimensionality of data represented using a visualisation is usually three dimensions or greater. Even when many well know methods exist for two dimensional or low dimensional data, visualisation works best when applied to data of higher dimensions.
- Visualisation concerns itself with data transformation (the continuous creation and modification of the meaning of the data) for better visualisation.
- Visualisation is in itself naturally interactive including its ability to be created, transformed and viewed.

Moreover, due to its focus on representing high dimensional information, visualization pipelines mainly focus on the data source and how operations are applied to transform it before representation. The Figure 2.4 illustrates the specific operations applied to data in a visualization pipeline.

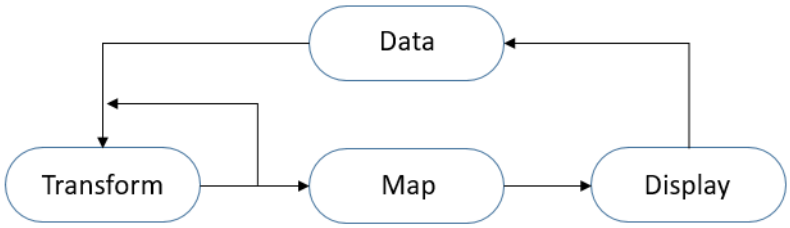


Figure 2.4: The visualisation process on how the data is transformed before the viewing it on a visual medium (Schroeder et al. 1998).

In the first step, the data is acquired using a sensor source. The data acquisition step is followed by data transformation. In this step, the data is transformed by various methods and further linked to an appropriate form of presentation in the mapping step. Finally the mapped data is then rendered or displayed on to a visual medium like computer screen or 3D display. Often the different process of the visualisation pipeline is repeated as and when data is newly available. This helps to better create accurately designed visualisations of the data.

To summarize, data visualisation is an powerful method to represent important information or phenomena. Effective visualisation approaches go a long way to make data understandable and explainable. Having understood the data, analysts and other decision makers can then make more informed decisions using it.

2.2.1 Data Representations

To design representations, we need to know specific properties about the data itself and understand how the property (attribute) change over time. For example, if the idea would be to visualise favourable routes on a map, then different properties that would quantify favourable routes (e.g., traffic congestion, air pollution e.t.c.,) have to be understood to create useful models and powerful visualisations. Inadequate knowledge about both the data and its attributes would otherwise result in inflexible or poorly designed visualisations.

However the understanding of a visualisation would involve an interplay between the visually represented metaphors and the users knowledge of the representations. This would in effect mean that it depends on what data is presented to users and how the user perceives the visualisation based on his/her domain knowledge. Some representations in GIS applications for example, focus on applying property specific information about the data by encoding it in the visual design (Roth 2017). The use of abstract, manipulable signs referred to as *visual variables* are frequently employed in map designs and cartography.

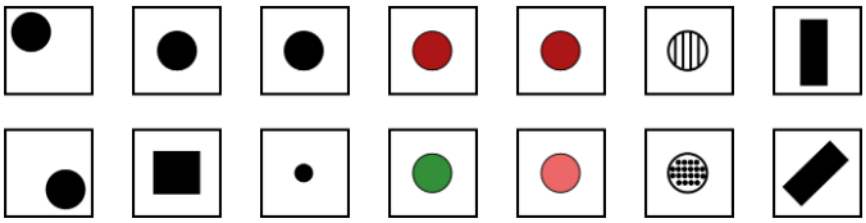


Figure 2.5: From left to right with each column showing a variant : Visual variables - position, shape, size, hue, value, texture and orientation.

Originally proposed by Bertin Jacques 1983, the basic set of visual variables used when designing maps include the following:

1. Position: Location based description of a map symbol relative to a coordinate system.

2. Size: refers to the amount of space a cartographic symbol occupies. Larger map symbols correspond to an increase in the value of the represented attribute.

3. Hue: refers to the predominant wavelength of the map symbol within the visible segment of the electromagnetic spectrum

4. Shape: describes the external form of a map symbol

5. Value: alteration in relative lightness or darkness

6. Orientation: alteration in the alignment

7. Texture: Alteration in Boundary sharpness

To enhance visual representations in map, the above mentioned list was then further expanded by other researchers including Joel Morrision (Morrison 1977) and Alan MacEchren (MacEachren 2004) to include variables - *Color saturation, Arrangement, Crispness, Resolution and Transparency.* These variables have since then played a crucial role in the design of map making. Amongst the different visual variables, crispness has been used to graphically represent uncertainty phenomena in geographic maps. The Figure 2.6 shows an illustration from (Fuest et al. 2023) of how favourable routes have been encoded using visual variables. The navigation paths choices shown are visually encoded with congestion information in all of the representations. To reach a given destination, if one is free to choice path A or B in either of the maps (a-d), then the choice made would not be the shorted path, but a route that could be most convincing even after the application of the variables to encode phenomena like traffic congestion along each of the routes.

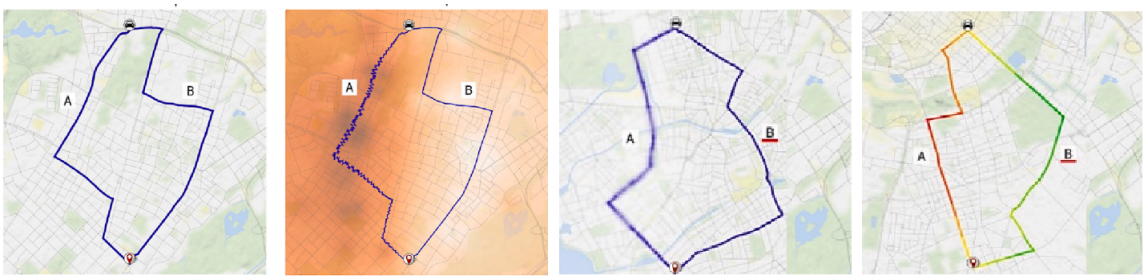


Figure 2.6: Visual representation to influence navigation choices of cars along driving paths A or B using (a) no visual variables, (b) line distortion along path A, (c) blur of path A (d) color coding the favourable paths B in (Fuest et al. 2023).

2.2.2 Visualising 3D Content

3D visualization is the computerized process of generating realistic and highly detailed digital content for three dimensional spaces. The use of 3D visualisation strengthens the visual representation of dynamism- *representation of motion* by using technology. Dynamic visualisations in 3D offers the opportunity to represent temporal process and complex dynamic phenomena that would take place in the real world. Such visualisations are being increasingly used in a number of application areas involving scene movement and data interpretation.

Based on the principles on which scientific content is created and displayed in geospatial context, 3D visualizations have been categorised (Seipel 2013) as either **weak** or **real** content. The term *weak 3D visualisation* is conceptually based on the idea that 3D models are projected onto a two dimensional display surface to view content. As content presentation of 3D onto flat surfaces involves projections, they are often prone to distortions. In such cases, the only clues for the observer are pictorial clues that would comprise shading, occlusion and changes due to the different sizes (Kjellin et al. 2008). Such visualisations are often also referred to as pseudo 3D visualisations. *Real 3D visualizations* engage both monocular and binocular depth cues (especially binocular disparity cues) to achieve stereoscopic vision of objects. Such 3D representations can then be displayed through different forms of visualizations and technologies. Examples of this class of display variants are monoscopic 3D display- desktop based 3D and stereoscopic 3D- AR or mixed reality.

2.2.3 Mixed Reality and Interfaces

Milgram et al. (1995) defined the Reality-Virtuality continuum as shown in Figure 2.7 and has been one of the early attempts towards defining Mixed Reality (MR). The continuum spans along two extreme environments (real and virtual) and contains real scenes on its left and a completely virtual world to its right. Everything that falls between the two extremes of real and virtual can then be classified as Mixed Reality (MR). Hence the types of MR as per this definitions can be either *Augmented Reality* (AR)- a real scene augmented with virtual content and *Augmented Virtuality* (AV)- when "either completely immersive, partially immersive or otherwise, to which some amount of (video or texture mapped) reality has been added". Also, AR is only a subset of MR as per this definition. It is important to note that these terms are however used often interchangeably in the scientific community and also in this thesis.

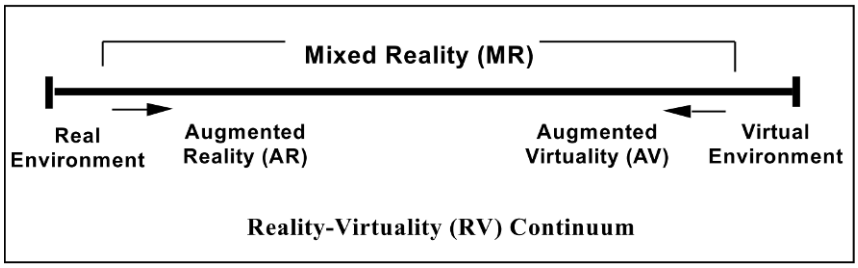


Figure 2.7: Reality-virtuality continuum (Photo: Milgram et al (Milgram et al. 1995)).

Based on the technical method of how AR experience is created and presented, interfaces in AR can be further classified differently. One approach, the **video see-through (VST)** is based on the concept of seeing the world through cameras. Typical examples of such interfaces include mobile phone-based AR and video see-through head mounted displays (VST-HMD). Such AR experiences come with the added advantage of creating visually coherent representation of information embedded directly onto the camera-image space (Collins et al. 2017, Kiyokawa 2016). In such interfaces, the user never sees the real world directly but through the videos feeds of cameras embedded to the devices. While this poses the inherent disadvantage of decoupling the user from reality, it also introduces other social issues with removing the ability to create eye contacts and social negotiations for headset users.

Another approach to creating AR experiences has been using **optical see-through** displays. Semi-transparent displays are used in the user's visual field to see superimposed computer generated graphics in this class of AR interfaces. The common examples include head-up displays (HUD) (Evans et al. 1989) and OST-HMD. In comparison to VST-HMD, optical see-through devices resemble closer to wearable glasses that makes them more acceptable for a wider range of applications.

The third form of augmentation is the use of projectors to augment the real world scene with virtual content. This is commonly referred to as **spatial augmented reality** or projection mapping (Bimber and Raskar 2005, Bimber et al. 2008) where visually coherent augmentation is projected onto flat surfaces or geometrically aligned uneven surfaces.

2.2.4 Visualisation Techniques in AR

As visualization of content for AR is three dimensional, the basic visualisation pipeline depicted in Figure 2.4 is modified to support 3D content placement with AR. When applying virtual content to an AR setting, the pipeline for it has to be modified to reflect not just the input data as mentioned in section 2.2, but a combination of real and virtual information while making the whole content presentation immersive. This would mean that the data source of the pipeline is not just raw data, but data appended with camera pose and scene image. The data to be visualised in AR (referred to as geometric data in Figure 2.8) is combined with AR inputs only in the compositing step. Until then the data of interest undergoes mapping/transformation like a normal scientific visualisation pipeline. Post the compositing stage using the transformed geometric and scene data source (image and registration) the AR view is constructed.

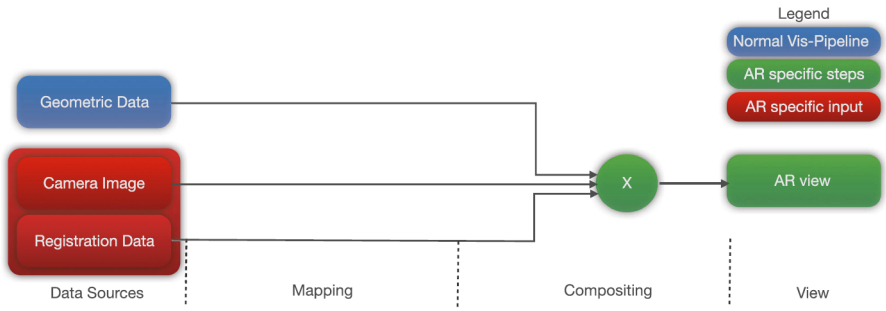
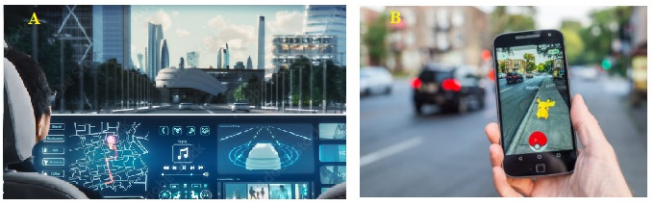
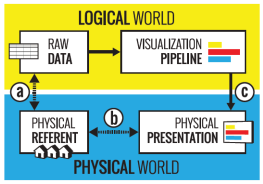


Figure 2.8: AR Visualisation Pipeline illustrating the modifications to a naive approach for AR content (Zollmann et al. 2020). A Depth camera will further enhance camera registration when included in the pipeline.

As the positioning of visual content is important for the pipeline, the resulting visualisations have a strong interdependence on where and how the content is placed in the input camera scene. As per the definitions of Zollmann et al. (2020), either a *situated* or an *embedded* visualisation could result from the output of an AR pipeline. The key difference is on how the virtual content is viewed with respect to the **physical referent**- the space or object associated with the data. The visualisation could be as virtual circles around cars when seen by an observer in the street or a 3D dashboard display with nearby cars appearing and disappearing based on their proximity as seen for a Tesla.



(a) Left - Situated visualisation inside a vehicle, Right - Virtual content embedded within viewing space.



(b) Visualisation pipeline with the data and presentations both linked to physical referent space (Willett et al. 2016).

Situated visualisation: In this data representation, the physical representation of information is located close to the data's physical referent but not placed directly on it. The visualisations belonging to this class are dashboard indicators for speed and real-time traffic status on a smartphone map.

Embedded visualisation: Overlays and projections of virtual content happens directly onto the physical referents in the case of embedded visualisation. Examples from this class of representations include existing systems that use see-through displays, projections and other augmented reality (AR) technologies to overlay the visualisation. This helps to see the content overlayed in the scene. This could be more appropriate visualisation choice than the situated case where the content is placed in a relevant location but not necessarily aligned to the corresponding data source. Hence AR visualisations should be designed based on the use cases and on how they are required to influence decision making when viewing the content.

2.2.5 Hololens and Mixed Reality Toolkit

The Hololens 2 (Figure 2.10) released by Microsoft is a head-mounted mixed reality device (OST-HMD) that has been widely used in both augmented reality and mixed reality. In 2018, Microsoft also released the research mode features along with the mixed reality APIs. This has allowed access to all its raw sensor streams (RGB camera, Depth, IMU, sound and microphone) for research purposes. With this, computer vision algorithms can be applied to the raw data to process and interpret the surrounding scene while using the Microsoft tools like MRTK to visualise virtual content.

Figure 2.10: Hololens 2 Mixed Reality Headset.

Mixed Reality Toolkit (MRTK) is a Microsoft-driven open-source project to accelerate cross-platform mixed reality development. The toolkit currently in its third generation (at the time of writing the thesis) supports Unity development using C# programming and is widely used by researchers and the scientific community. The MRTK is designed to be a quick access toolkit for developers to create high-quality and performance-focused virtual experiences. Designed for performance for the Hololens 2, it is optimised for other resource-constrained mobile platforms and supports different interaction paradigms (touch, input and audio). Also, it is supported by OpenXR- an open-source royalty free standard for virtual reality and augmented reality. The MRTK software toolkit enables developers to build applications not only specific to the Hololens, but across a wide variety of devices including Meta Quest, SteamVR, Oculus Rift and Lenovo ThinkReality. The toolkit provides a range of features that include:

Hand Tracking: This feature supports visualising and representing the hands of the Hololens user in a mixed reality. This allows for a more natural interaction compared to the use of controllers (in VR e.t.c.,) but often could result in a lower precision for hand movement tasks like drawing or writing. Each finger, joint and the thumb are represented using three and two points (Figure 2.11) in the virtual experience. Another mesh based representation for better performance is also supported wherein the hands are visualised using a mesh material (Figure 2.11).

Eye Tracking: The toolkit provides developers with ability of understanding what the users of the Hololens are looking at. The privacy-preserving eye tracking API avoids passing any

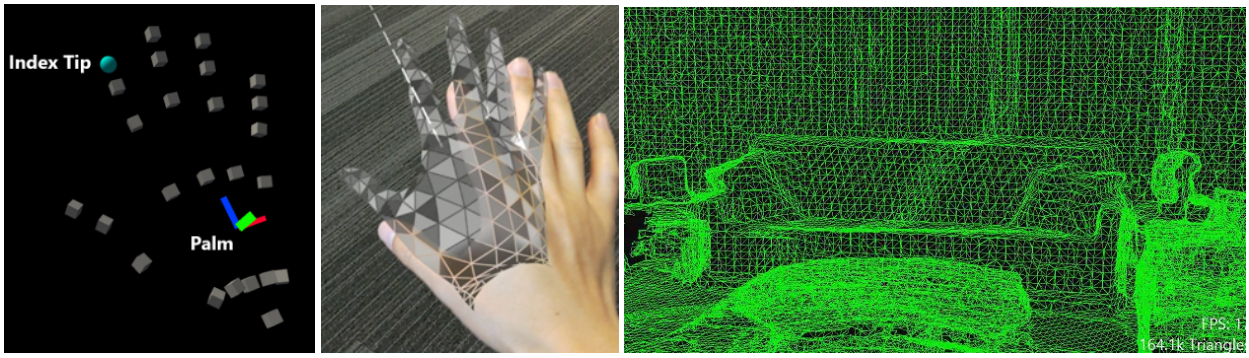


Figure 2.11: The figure shows the different features supported by the Microsoft MRTK with Hololens hand tracking using joint model (left), mesh model (center) and the spatial mapping feature of the toolkit creating model models of indoor scenes (Microsoft 2023).

biometric information and captures gaze data from the left and right eye simultaneously. To improve the accuracy of the tracking, the user is required to first calibrate the eye by looking at a set of holographic targets. Post successful calibration, the API provides information on where the Hololens is looking at with the eye gaze ray (gaze origin and direction) at 30 Hz

Spatial Mapping: The spatial awareness feature of the toolkit constantly tracks the environment using the RGBD sensors, creating a 3D model of the area around the headset. The depth cues acquired from the Hololens build in sensors are then used to add occlusions and depth effects to the 3D objects placed in the view of the user. These effects applied to the 3D content (in real-time) create convincing user effects allowing to bend virtual objects to the surrounding scene.

2.3 Depth Sensing and 3D Scene Data

2.3.1 Time of Flight Sensing

A Time of Flight (TOF) sensor is a type of distance measurement device that uses light or electromagnetic waves to determine the distance between the sensor and an object. The sensing technology is a high framerate 3D imaging technique producing intensity images and range data for every image pixel acquired using the sensor. The time needed by the light emitter to travel from its source and reach the receiver is proportional to the distance of the sensor from reflected objects. The ToF depth is measured by using either **Pulse** or **Phase** modulation as shown in Figure 2.12a and 2.12b.

Pulse Modulation: In this distance measurement approach, the round-trip time it took for the light pulse to return is measured. The round-trip time of the pulse and speed are then used to estimate the distance using Equation 2.1. Moreover, for a higher distance measurement accuracy, very good cock circuits are required in this approach.

$$d = \frac{t}{2} \cdot c \tag{2.1}$$

(a) Pulse modulation.

(b) Phase modulation.

Figure 2.12: Time of Flight (ToF) Sensing.

where d is the distance between the sensor and object, c is the speed of light and t is the time between emitted and received light source.

Phase Modulation: For this method of ToF distance measurement, the emitted light from the sensor is modulated using a signal modulation technique (Sinusoidal, FMCW, pesudo noise or polarisation). The receiver then measures the phase shift of the received signal.

2.3.2 Pinhole Camera Model

Most ToF sensors available commercially are used for depth measurement and are part of RGBD devices. These RGBD devices have both an image camera and a depth sensor that output an RGB image along with its 2D depth map. Also as it might be hard to observe finer details of depth from just a 2D depth, the camera pinhole model (Figure 2.13) is used to re-project the depth onto the 3D space.

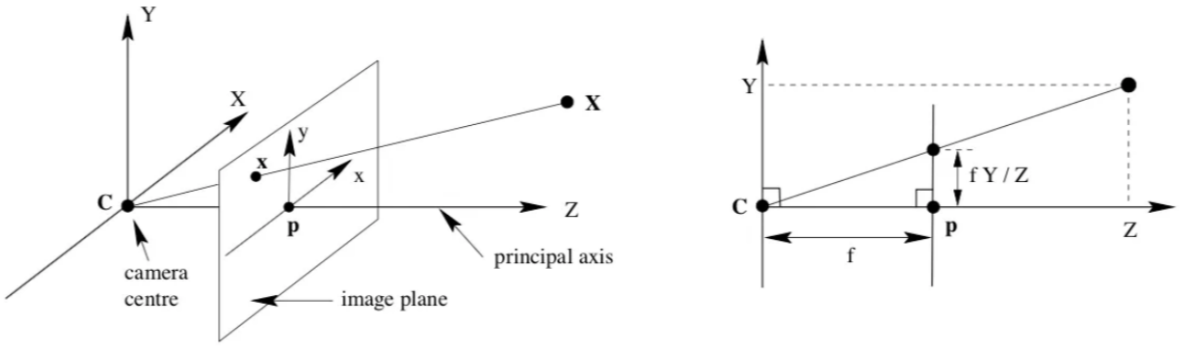


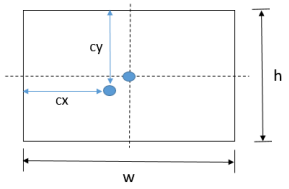
Figure 2.13: Pinhole camera model.

The pinhole model describes the mathematical relationship between 3D points of a scene and its corresponding 2D points captured from a camera perspective. The intrinsic matrix **K** (Equation 2.2) of the camera transforms 2D image coordinates to 3D and vice versa. Hence each 3D point is mapped to 2D pixels (u and v) by applying a transformation (u, v) = f(X, Y, Z)

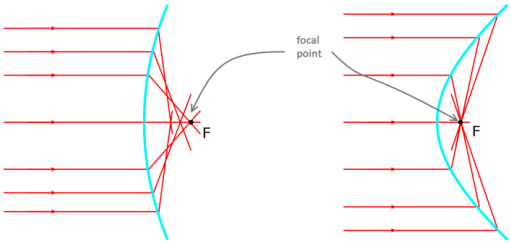
$$K = \begin{pmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

Along with the camera intrinsics K stated above, other distortions of the lens and extrinsics transformations have to be accounted for while applying computations. The complete mathematical model that describes the transformation of the image point p to its corresponding 3D point P can be written as $p = K[R \mid t] * P$. Where K is the camera intrisincs and R and t are the extrinsics that are further detailed in the next sections.

Intrinsic Parameters (K): Each intrinsic parameter describes a geometric property of the camera. The internal geometry and its optical properties are best described using the focal length, principal point and lens distortion. The focal length (f_x, f_y)- the distance between the optical lens center and camera sensor, the principal point (c_x, c_y)- the displacement of the optical axis from the camera projection center and the distortion coefficients- the inherent optical distortion of the camera; best describe the intrinsic properties of a 2D camera source.



(a) Camera principal point.



(b) Different projections for spherical (left) vs parabolic lens (right) (source: Villena-Martínez et al. (2017)).

Figure 2.14: Figure on the left depicts the optical center while right shows the camera distortions.

The distortion represents the variations in light projections due to optical aberrations in the camera lens. The amount of distortion is highly dependent on the position and distance of the principal point. This error will be zero at the principal point and increases with increasing distance from the point. The optical distortions are best represented using the distortion coefficients in the camera model. The two common distortions are:

Radial Distortion: This type of distortion is the most visible one affecting low-grade camera and lens. The distortion occurs due to the unequal bending of light with rays (Figure 2.14 (b)) bending more near the edges and lesser near the centre. Hence the distortion gets more pronounced as we move away radially from the optical center.

Tangential Distortion: This occurs when either the image screen or the sensor is not parallel to the lens but at an angle w.r.t the lens. This distortion would result in objects appearing farther away or closer than they are. Also, images in most cases would look titled or stretched on viewing.

Extrinsic Parameters (R, t): The extrinsic calibration between the RGB image and the ToF depth sensor refers to the 3D geometric relationship to register information from differently positioned RGB and depth sensor rigidly mounted on the same device. The extrinsic transformation is defined by a rotation (R) and translation matrix (t) which represents the orientation and displacement between the image and ToF sensors. By applying this transformation, it would be possible to register depth information in every RGB image pixel.

Depth Error and Compensation: The Depth measurements from an RGBD sensor are often prone to both systematic and non-systematic errors. In general systematic errors are caused due to the intrinsic properties of the depth camera or due to imaging conditions prevailing when capturing the data. As these errors are relatively fixed, they can be evaluated in advance and have corrections implemented. Of the two errors, the systematic errors can be reduced by good calibration procedures. Amongst the systematic errors, depth distortion is the most prominent error source that affects depth measurements.

Depth Distortion also known as wiggling / circular error arises due to irregularities in the modulation process when capturing depth. This distortion occurs when the emitted light from a ToF sensor is not generated mathematically modulated with the computed modulation (e.g., sinusoidal modulation in the case of amplitude modulation). Such errors can be corrected by using different calibration techniques that would then include calibration targets and chessboard patterns.

2.3.3 Hololens Research Mode Sensors

The Research Mode has promoted the use of Hololens as a powerful tool for doing research in computer vision and robotics while also using it for visualising 3D content. HoloLens 2 (Fig. 2.10), which was announced in 2019, brings several improvements with respect to the first-generation Hololens device– like a dedicated DNN core, articulated hand tracking and eye gaze tracking.

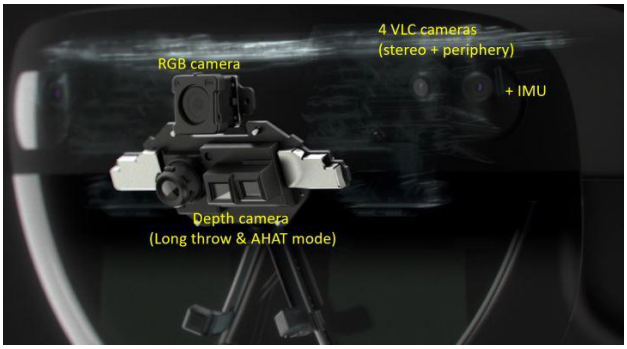


Figure 2.15: Hololens 2 research mode sensor streams.

The second generation AR device features a custom-built Holographic Processing Unit (HPU 2.0), which enables low-power, real-time computer vision. The HPU runs all the computer vision algorithms on the device (head tracking, hand tracking, eye gaze tracking, spatial mapping etc.) and hosts the DNN core. The device is equipped with an array of sensors to localise itself using SLAM as shown in Fig. 2.15. The specifications of these sensors

are detailed below. Also, the research mode allows sensor access using public repositories containing tools and sample applications (Ungureanu et al. 2020).

Research Mode for HoloLens 2 enables access to the following input streams:

- Four visible-light tracking cameras (VLC): Grayscale cameras (30 fps) used by the system for real-time visual-inertial SLAM.
- A depth camera that works on the principle of ToF sensing and operated in two modes:
  - Articulated Hand Tracking (AHAT)(45 fps), near-depth sensing used for hand tracking. As hands are supported up to 1 meter from the device, the HoloLens 2 saves power by calculating only "aliased depth" from the phase-based time of flight camera. This means that the signal contains only the fractional part of the distance from the device when expressed in meters.
  - Long Throw (1fps), far-depth sensing used for spatial mapping on device. This sensor can be used to detect objects in the scene including moving pedestrians.

  Both AHAT and Long Throw are also supported by the Active Brightness (AB in short) feature as mentioned in Table 2.2. The depth sensors captures data in its own local cordinate system.
- Two depth modes of the IR stream (AB), are computed from the same modulated IR signal for depth computation. These images are illuminated by infrared and unaffected by ambient visible light.
- Inertial Measurement Unit (IMU):
  - Accelerometer, used by the system to determine the linear acceleration along the x, y and z axes as well as gravity.
  - Gyroscope, used by the system to determine rotations.
  - Magnetometer, used by the system for absolute orientation estimation.

Stream	Resolution	Format
VLC	640 x 480	8-bit
Long throw depth	320 x 288	16-bit
Long throw AB	320 x 288	16-bit
AHAT	512 x 512	16-bit
AHAT AB	512 x 512	16-bit

Table 2.2: HoloLens research mode sensor resolution and format.

2.4 Object Detection And Tracking

The goal of an object detection algorithm is to detect all instances of an object (e.g., pedestrians, cyclists and vehicles) from one or several known classes as in the KITTI dataset (Geiger et al. 2012). The input to the algorithm in each case will be an RGB image or its 3D depth. Each detection is then reported with an associated *pose* information and a probability score. The pose could then be (a) a position of where the object is; or (b)

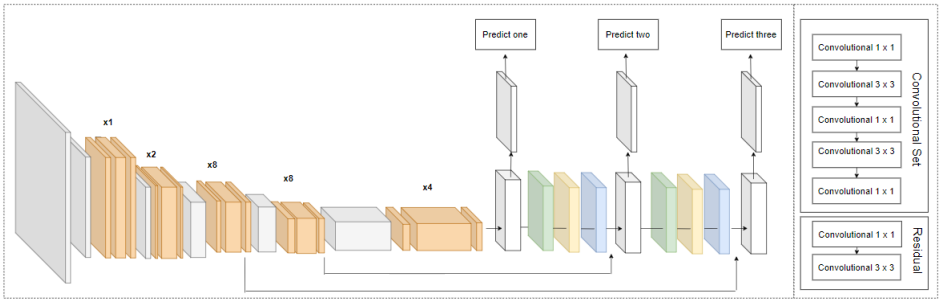


Figure 2.16: YOLO network architecture (Source: Mao et al. 2019).

an enclosing bounding box indicating the object presence or (c) a segmentation mask that differentiates the object for all other objects in the scene.

Object tracking on the other hand tasks itself with inputting an initial set of object detections, creating and maintaining a unique ID for each object; and tracking objects as they move around in the frames. The IDs that were assigned are managed by the tracker during this process.

2.4.1 Image Only Object Detection

Visual object detection aims to find objects of a certain class with precise localisation for a give image while associating each object with its corresponding label. Each object is then predicted with a bounding box center in pixel coordinates (cx, cy) and box dimensions (w, h) that would enclose the object. A more richer understanding of the scene is provided by *semantic segmentation* that predicts a pixel-wise classifier mapping each pixel to a specific label.

Deep learning based detectors that are widely used for object detection can be primarily divided into two: (i) one-stage detectors and (ii) two-stage detectors. In the case of YOLO, a **one stage** detector, there is only a single neural network evaluation for object inferences. For this the model will use pre-defined set of boxes (anchor boxes) to look for objects. In terms of speed, the inferences from such networks are quick to run in real time (high frame rates) and hence support augmented reality use cases. The **two stage** algorithms are more accurate but also more complicated. The scene is first divided into multiple regions in the region proposal step. After screening out the negatives, the regions of interest (ROI) are generated. This is further followed by region classification and location refinement of the ROI. The common examples of two stage detectors include Faster R-CNN, R-FCN, FPN, etc.

YOLO: You Only Look Once (YOLO) is a widely used object detection method that was developed by Redmon et al. (2016) and improved in performance for the multiple versions that followed. With the YOLO, object detection is looked upon as a regression problem in which the objects in the scene are determined from the pixels of the input image using a neural network. An object recognised by the network is output as a 2D bounding box which then indicates the position and size of the object in the camera scene. A class probability is also estimated for each predicted object. All objects recognised by the network in the scene are assigned higher probability scores.

Compared to region proposal networks (Fast RCNN) which perform detection's on various region proposals and end up performing multiple predictions for various regions of the same image, Yolo like a CNN passes the image through the network once and outputs the object bounding box predictions. The input image is divided into a grid and to each grid cell is set of anchor boxes that are associated with the same centroid. The approach then computes how much of the ground truth box overlaps with the anchor box picking the one with the largest overlap. This is followed by the prediction step where offsets to the anchor boxes are then predicted.

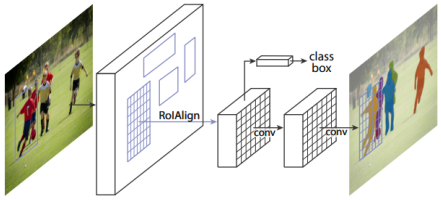
Among its predecessors, Yolo v3 can detect objects with higher accuracy even when they are close by the camera. The Darknet-53 which contains 53 layers is used as the backbone for Yolo v3 with certain outputs forwarded back to the network as in the ResNet architecture (Targ et al. 2016). The total number of stacked layers in YoloV3 (Figure 2.16) total upto 106 layers with the network outputting each detected object class, its bounding box and a probability score.

The biggest advantages of YOLO when comparing to other image detectors are :

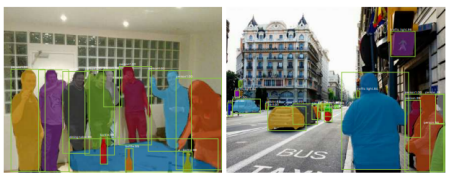
- Speed and faster inference
- Network understands generalised object representation
- Faster with a smaller architecture and open-source

MaskRCNN: MaskR-CNN introduced by He et al. (2017) is one of the most widely used image based DCNN networks that detects objects with a bounding box, image mask and a semantic class. The network simultaneously solves both object detection and instance segmentation for each of the objects for a given input image.

Mask RCNN holds architectural similarity with the Faster-RCNN (Ren et al. 2015) network, a predecessor object detection approach. While the latter has two outputs for each candidate object - a class label and a bounding box offset, MaskRCNN has an additional branch that also outputs the object semantic mask. Hence the network uses region proposal, object classification & box regression and instance mask segmentation as a part of its pipeline to make a semantic inference about input scene objects.



(a) MaskRCNN detection network architecture



(b) MaskRCNN pedestrian detection on COCO test sets (Source: He et al. 2017).

Figure 2.17: MaskRCNN semantic segmentation for RGB images

Firstly, a ResNet 101 backbone model is used to extract features from the images which is then passed to the region proposal branch. In the next step, the Region Proposal Network (RPN) predicts if an object is present in the region or not making an inference of the existence of objects in the feature maps. As the outputs from the RPN would be of different shapes, a pooling operation is applied converting all regions to the same shape. In the final step, the network branches out into two:

– **Object classification and regression:** Here the network predicts the class and position of objects using the obtained regions from the previous step.

– **Mask segmentation:** This branch employs a fully convolutional network to produce k instance binary masks m x m , one for each proposal regions. The mask that matches the predicted object type is then scaled up to match the size of region of interest (RoI).

The semantic segmentation network has a proven inference time of 195ms (per image) on an Nvidia Tesla M40 GPU (He et al. 2017). The network achieves good results even under challenging conditions (Figure 2.17b) and is designed to estimate human poses with minor modifications.

2.4.2 3D Detection using RGBD Sensor Data

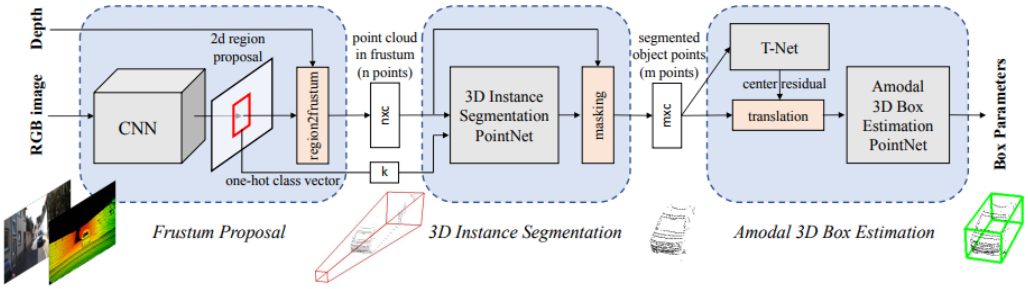


Figure 2.18: Frustum Pointnet for 3D object detection.

The trend towards applying both images and depth to detect objects in 3D has been the focus of many RGBD based object detection algorithms. In most of these approaches, the works have focused on applying depth data to a proven image detection CNN to localise objects in 3D in the subsequent stages. For example, based on RGBD data, Kollmitz et al. (2019) extended a Faster R-CNN model to regress the 3D centroids of pedestrians. Meanwhile Linder et al. (2020) extended the YOLO v3 model to directly regress the 3D centroids. In addition to the regression of 3D centroids, Explanable YOLO (Takahashi et al. 2020) used a 4-channel RGBD data directly as input to regress the 3D bounding box of pedestrians using Darknet-53 backbone network.

Along with image-fused methods, object detection using Frustum-based Pointnets (Qi et al. 2018) have proven to be used for real-time pedestrian detection (Shenoi et al. 2020). In the F-Pointnet approach (Figure 2.18), objects are first detected in 2D images which is used to create Frustum point clouds. Then the foreground points and its features are extracted using the Pointnet network (Qi et al. 2017). These foreground points are then used to estimate the 3D bounding boxes of objects in the scene. Also the method has been proven to work well for indoor scenes and brightly lit outdoor scenes.

The performance of F-Pointnet could however be limited:

– based on the object predictions of the image detector used in the frustum proposal

– if very few points are identified as foreground points to regress a 3D bounding box.

2.4.3 Particle Filter Tracking

Object tracking is the process of following a particular object or multiple objects (Djuric et al. 2003; Cappé et al. 2007) in a sequence of frames for a video or 3D scene. The goal of tracking is to follow an object of interest over a data sequence by monitoring its movement over time while dealing with issues like occlusions, appearance changes due to illumination differences and scale issues. Also based on the number of objects that are tracked , they can be classified as Single object tracking (SOT) and multiple-object tracking (MOT). While single-object tracking would estimate the trajectory of a target object over time for a given initial location in the first frame, MOT would track multiple objects in the scene simultaneously.

Particle filters is a widely used tracker designed on the bayesian principle. The filter applies bayesian equations to solve for the motion attributes (e.g,, position velocity) of the objects and track their state. This means that the probability estimate of the state at different time instances can be formulated as a combination of *smoothing*, *prediction* and *update* in the filter.

$$p(kt_n|d_{1:t}); 1 > n > t - smoothing \tag{2.3}$$

$$p(k_t|d_{1:t-1}) \quad - prediction \tag{2.4}$$

$$p(k_{t+n}|d_{1:t}); n > 0 \quad - update \tag{2.5}$$

p(k_t) is the posterior probability of the state vector k_t and $d_{i,t}$ the respective observation vector in Equation 2.3, 2.4 and 2.5.

In the filter, current states of tracked objects are sequentially approximated (*smoothing*) and the model is updated with new observations/sensor measurements upon positive association (*update*) of each observation to the state. The unassociated tracks don't update and the final estimation is based on the motion model (*prediction*). Hence each tracked object can be formulated using the equations:

$$p(k_t|k_{t-1}) : k_t = f(k_{t-1}, \xi_{t-1}) \tag{2.6}$$

$$p(d_t|k_t) : d_t = h(k_t, \eta_t) \tag{2.7}$$

f(.) is the prediction model, h(.) demotes the sensor model. ξ_{t-1} is the system noise and η_t is the measurement noise in Equation 2.6 and 2.7. The objective of tracking is then to estimate the optimal state k_t that maximizing the belief given all the past observations $d_{1:t}$. Hence the Bayesian inference allows for estimating a state by combining a statistical model for a measurement (likelihood) with a prior probability using Bayes' theorem.

Particle filters use Monte Carlo simulation to represent the posterior PDF as a weighted sum of discrete samples called particles as represented by the Equation 2.8.

$$p(k_t|d_{1:t}) = \sum_{i=1}^{N} w_t^i \delta(k_t - k_t^i) \tag{2.8}$$

k_t^i is a random sample, δ is the Dirac delta function and w_t^i are sample weights.

For a given observation d_t, the likelihood $p(d_t|k_t^i)$ can be computed from the observation model defined in the equations stated above. Hence with the likelihood of the samples, an approximation of the state is estimated. The complete set of particles then represent probability function, such that the final estimated state $\widehat{k}$ can be obtained by searching for the mode of the distribution. The tracker can then be used to track objects like pedestrians, cyclists and cars based on the choice of the motion model.

2.5 User Study Design and Considerations

User studies offer a scientific and a proven method to evaluate the performance of visualisations. While developing successful visualisations, their disadvantages and merits have to be clearly understood and hence the importance of evaluating them accurately cannot be overlooked.

The start point for a use study is a general research question that needs to be addressed with the visualisation. The research question that needs to be addressed is the core component for a successful experiment design. When the question is well refined, the methodology that would best suit the evaluation can be selected to answer the questions. A clear question is important as this would clearly state what has to be investigated and how the research can be carried out to collect data for analysis. In addition to formulation of the question, the choice of task would require knowing what questions can be addressed within the scope of the study.

2.5.1 Designing an Experiment

Once its clear on what is known and what actions have to be possibly taken, we can run a set of tests applying visual variations with each experiment execution. Each single execution of a variation is called a *trial*. The full collection of trails that addresses the research question is referred to as an *experiment*. For example, if the objective of a study is to test different versions of visualising rivers in a geographic map, then different variants of colours are tested in the different trails of the experiment to study river visualisations.

As visualisation is a perceptual task, we cannot directly measure the influences or/and impact of each trial completed. Therefore we try to infer of how people reacted to the information they saw. Then, by observing both the inputs (visualisations) and outputs (reactions) we make an interpretation and a following inference. If we let $\mathbf{x}$ be the vector representing description of the situation and M(x) the measured response, B(x) could then represent the internal process of a *perception-action-loop* that interests us to make inferences.

$$M(x) = B(x) + e_w \tag{2.9}$$

The error term (e_w) in the equation 2.9 signifies the deviation between the performed reaction by the participant and the reaction he/she intended to perform. The term represents the unintended variation in human behaviour in repeating tasks. One cannot produce the same action twice, no matter how hard one tries. Even when e_w is different with every measurement M(x), B(x) is assumed to be a constant. Most evidences however suggests that B(x) would remain constant *for short periods of time*. This has been one of the two reasons to keep experiments short. Fatigue of participants has been another reason for short

experiments. Furthermore averaging several measurements would yield a good estimate of B(x). This would in the process increase the likelihood of averaging out the error term e_w from the equation.

As for when trails are repeated with the same participant in an experiment, the second trail would differ from the first as participants would have had more practise with the second attempt. Hence perception of subsequent stimuli and their performance in the task could be impacted by what has already been seen earlier. This would mean that not controlling for order has introduced a confusion. The influence of repeated ordering can be reduced using one of the three design approaches as in Figure 2.19:

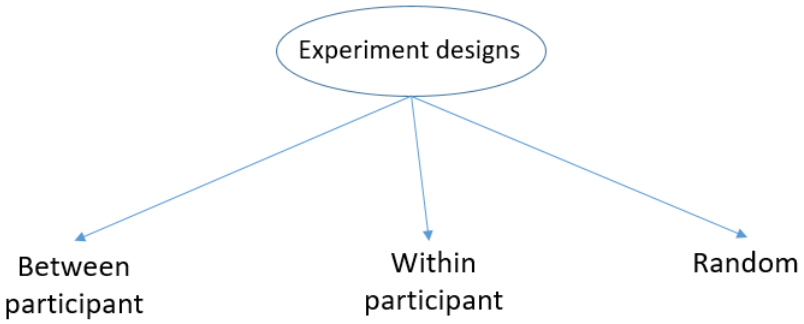


Figure 2.19: Different designs for a user study experiment.

Between participant design: In such designs a single person sees one and only one condition. As one person sees only the condition once, the effect of ordering is removed. Furthermore as different participants see the same condition now, other effects due to inter-person differences come into effect. The measured difference between conditions would now reflect a compound effect of both the stimulus and participants respectively. For example, in a visualisation study focused on how people understand different colors, if one person is shown red and the other a green, then the difference in performance is due to both from differing stimulus or due to intrinsic difference between the two participants in the study. Such a design is based om the assumption that the between participants error is small.

Within participant design with ordering variations: The fundamental idea here is that between participant effects would not show up if all measurements are made with same person. This would then mean that in such case, every person sees all conditions and then the person would act as a baseline himself or herself. Moreover to remove the effects of ordering, different groups would see different orders. This would mean for the same visualisation study stated in the previous point, one group will have participants seeing the red first and then followed by the green; while another group are shown the two colors in the opposite order (green followed by red). A fundamental issues with such designs would be the possible orders to be examined. This would mean for three conditions there could be six possible orders, which would then be 24 possible orders when the conditions to be tested are four. Which might turn out to be expensive when the possible conditions become X (X! orders)

Random ordering: With this approach every participants sees the stimuli in randomly chosen order. This means that samples are collectively averaging the effects of noise due to

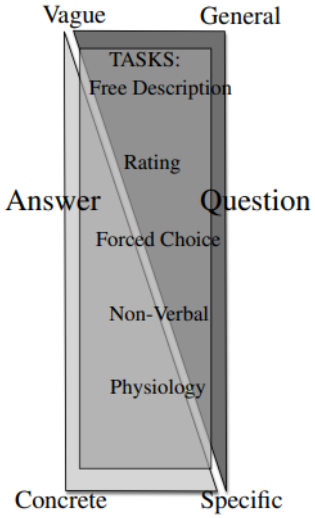


Figure 2.20: Task continuum proposed in Cunningham and Wallraven 2013.

ordering. Another way of inferring this would be to a sampling approach from many order groups via fully controlled order experiments. This also means that such an approach needs more participants.

2.5.2 Performing an Experiment

Once the experiment has been well designed and planned, the subsequent steps would include conducting the study using a sample population. The following elements constitute a successful experiment:

Participants: Each experiment is focused towards a target population. The choice of the population is highly depend on the skills and domain knowledge needed to understand the tasks for each of the experiment trails. Once the target group is fixed, it is essential to ensure that there is enough random sampling of the population to have results that could generalise and be representative of the population.

Tasks: While there are astonishingly large number of tasks in behavioural science research, a chosen task for an experiment is designed to answer specific types of questions. Quiet often the task is implicit in the question that is asked. Tasks and its complexity considered for an experiment can explained using the task-continuum (Figure 2.20) proposed in (Cunningham and Wallraven 2013). On the "general" end of the continuum are the meta-tasks of how the person would react for a given situation. This is done by sharing questionnaires where using ratings or free description text that would help to understand the objective of the study are captured. On the "specific" end of the task continuum are the tasks that can concretely estimate inferences for very specific questions. This would help to map what elements of the visual stimuli the participants saw and how they inferred the presented stimuli.

Stimulus selection: The chosen visual stimulus forms the input to participants for the experiment. This visual input should be presented in a controlled manner using either a 2D

or 3D visual interface. Also the chosen visual method should be relevant for the task and the research question that is addressed.

Stimulus presentation: The medium of communicating the stimuli during the experiment is fundamentally an important component. The medium in this case will act as a channel in communicating the stimulus clearly to the participants.

Analysis: This step follows the data collection stage where the data of the responses of the participants is first collected before analysis. In the analysis stage, appropriate tools and methods are applied to the data to interpret how the participants reacted to the presented stimulus in the different trails of the experiment.

The underlying principle of user study research is to investigate and compare participant responses under different experimental conditions. For this one (or more) variables are manipulated to observe the impact of the change and investigate the effect on one (or more) other factors. While the former in this case is the independent variable, the later would be the dependent variable. Statistical test are then applied to the data to investigate the experimental effects.

Answering a given research question would require that many experiments have to be performed often using different tasks. No single experiment can control for all possible variances in infinite dimensions. Also no single experiment can have infinite sampling that would in the end be interest for the question. The need for control for errors in the experiment would require that we plan for the experiments well ahead of in time. Some methods might compel towards promising quick answers, but might later turn harder to analyse . Other tasks might require more careful and considerate planning; but then would not just make experimentation easily, nevertheless would reduce efforts during the analysis phase and quicker results in the process. The tasks then would be more clearer with less possibilities for misinterpretations. Hence designing an experiment with care and consideration is vital for successful interpretation of the results.

2.5.3 Ethics, Privacy and Confidentiality

While designing experiments, ethics, privacy and confidentiality of the participant data captured during the experiment have to be kept in mind. Also performing experiments in every country require approvals from the ethics review boards. Ethics review boards ensure participants and their data would be protected. This has been a reason that many journals do not publish results unless proper ethics guidelines have been followed in the work. Failure to follow ethical guidelines or approvals could be considered scientific misconduct and could lead to serious and legal repercussions.

Some people fear that the information that they provide in a user study could be used inappropriately. Many techniques exist for protecting each persons privacy and ensuring that the data shared would be used for only the clearly stated purpose. Hence each survey study needs to reassure the respondents that data privacy techniques have been incorporated. Also, the participants need to be convinced that the data would be collected anonymously and that no one will be able to map the collected data to their personal information.

The term confidentiality refers to the safeguarding of information about one person that is known to the other. A surveyor with names and addresses of people even in code should not use this information to reveal the identities of study participants. Confidentiality especially is a serious concern in online studies as even an email address or Internet Service Provider (ISP) address leak might lead to the survey respondent's identity being revealed.

3 Related Work

3.1 Walking Influences

This thesis proposes new ways to influence the walking behaviour of pedestrians by using Augmented Reality. Hence the literature study for the work follows a taxonomy distinguishing motion influences applied in current movement studies to the different levels as depicted in Figure 3.1. This taxonomy is partially borrowed from (Ishii et al. 2016) where visual influences to walking are classified based on how the input stimuli was shown and what movement influence resulted to the person. As per our taxonomy, influence studies can be distinguished to be either *Active*, *Passive* or *External*.

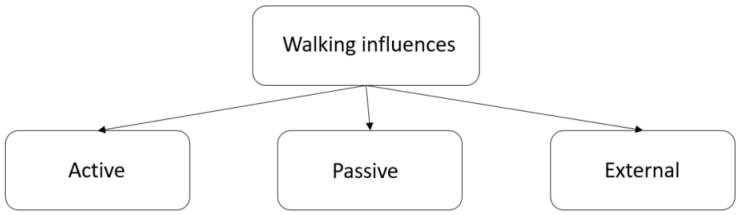


Figure 3.1: Methods to influence walking using visual influences.

Active Influences pertain to those studies and work where visual information is presented to the user with the intent of actively altering a person's walking path. This for example could happen when a person decides to change his/her walking direction via an active intervention. Such interventions could happen by using a smartphone showing a map or when he/her sees signboards indicating which path should be taken next. Passive on the other hand includes all those influences that subconsciously prompts a person to make an alternative path choice. The visual information seen could be either augmented on a visual medium to force him/her to take new paths. Lastly, external influences are all unaccounted factors that exist in a real world scene prompting people to unexpectedly move away from their intended motion paths. An example of such a external factor could be the sight of an approaching cyclist or a group of nearby pedestrians walking.

This chapter provides a brief updated review of the existing methods focusing mainly on the following aspects: (a) how is virtual content used to influence walking, (b) what medium was used to show the content visually to the user and (c) how was the presented information processed before the visual rendering step. By reviewing the existing methods, the knowledge gap that this thesis attempts to fill is highlighted in the last section.

3.1.1 Active Influences

The section reviews methods that emphasize actively communicating movement paths which influence walking decisions made while moving.

Recommendations Based on Past Movement Trajectories: A relatively quick and easy approach towards motivating people with AR to walk along a specific path has been by showing walked paths from the past data or preferred future walking directions. Projection based approaches have been researched where walking steps and trajectories were displayed on the floor to influence navigation decisions (Albarrak et al. 2019) and to create social awareness (Monastero and McGookin 2018). In a further study, the effectiveness of showing such guidance was tested with a virtual environment (Sakamoto et al. 2019). The environment for the study consisted of floor coverings that were illuminated to guide different pedestrians to the same destination. By projecting a specific pattern-constant flow of black and white strips on the walking surface, persons were guided to their destination. With such a dynamic visual approach, 7 of the 10 participants were influenced in the walking study. Extending such influences to public places, Albarrak et.al in (Albarrak et al. 2020) visualised the motion heatmap and footsteps as movement traces. For the visuals that were shown, other factors like personal preferences of paths and environmental constrains proved to have played a role in all the navigation decisions that were made by the people. Rather than focusing on historic movement data, another study (Albarrak et al. 2021) applied visualisations to an interactive floor display. In this work people in the scene were tracked and their footsteps were then projected on to the floors. Even in this work the way people walk with such visualisations highly depended on the way participants interpreted the visual clues shown on the floor.

Extending such visual design and influence methods for pedestrians in traffic situations has found itself limited to mainly concept ideas and basic prototypes. An AR based futuristic interface was prototyped in (Hesenius et al. 2018) to display walk-able paths for persons. The AR paths shown were then based on the traffic situations of the scene and tested in a study. Each person saw a virtual lane that was projected onto the real infrastructure where s/he could walk safely. However the visual variants tested in the study were mostly limited to only single pedestrian routes and were not scalable. The work did not address what would happen if there are more persons in the scene and how the virtual pedestrians lanes would expand or change to accommodate for the increasing number of persons in the scene. The work however noted that the success of such a visualisation would depend on the trust in the system for its correctness and trustworthiness.

AR Headset based Navigation: AR Navigation can be considered a special case of 3D content influenced walking. In most such navigation applications, the target destination is used as a prior and an AR overlay of the path is shown based on the shortest path computed (Grasset et al. 2011). Then the application user is influenced to walk along the shortest path to reach the destination based on his/her current position and the computing algorithm that estimated the proposed path. Even when such approaches are currently widespread using mobile devices (Papageorgiou et al. 2020, Shaheen et al. 2016), lesser works have applied 3D visualisations to navigation while walking with AR headsets.

The pioneers in navigating using headsets and applying 3D to it has been the earlier works from Feiner et al. (1997). In this work, the touring machine was prototyped to overlay 3D information onto the outside world to guide people to way-points. The content shown to the person via a body worn headset depended on the position of these persons and their proximity to real world objects. The prototype in the work combined position tracking using differential GPS and AR visualisation to achieve context based virtual content. Based on

the position of the person in an outdoor space, location specific virtual content was overlaid onto to a head mounted display. This was useful to get landmark specific information while walking. The work however did not support effective guidance to walk along specific paths. Another work (Höllerer et al. 1999) on the contrary, mainly focused on providing AR guidance to users by operators working remotely . In the MARS project that has been detailed above, motion guidance was provided via collaborative efforts based on teleoperation. Walking assistance to outdoor headset users was possible when a group of users stationed indoors sketched paths or pointed our interesting objects to them. While lesser works have focused on applying guidance via AR headsets, more recent works on AR and pedestrians have used smartphones (Dünser et al. 2012, Dey et al. 2018, Santos et al. 2014) and light projections (Colley et al. 2017b, Avila Soto and Funk 2018, Knierim et al. 2018). However the most important shortcoming noted in smartphone or AR navigation systems (Narzt et al. 2006) are the issues of visual attention. Most of the above mentioned applications use an arrow visualisation pointing the intended direction from an ego prospective or "Birds eye view" map. The visual interfaces then requires the user to pay attention continuously while in motion which could be distracting and attention demanding. The users are expected to both recognise the information presented and at the same time follow the instructions presented to reach the target goal destination.

3.1.2 Passive Influences

All forms of subconscious navigation methods that prompt persons to take a different walking paths than the consciously intended one are covered in this section.

Direct Manipulation of Visual Content: The visual interpretations of once own motion (self motion) is largely affected by what a person estimates the movement pace to be based on his/her sensory inputs from the surrounding. Also as the person walks, he becomes more focused towards his destination and starts to pay lesser attention to the surrounding. Visual information of landmarks and the resulting optic flow (Lappe et al. 1999) then start to dominate his/her vision.

Some motion studies have tried to manipulate the feeling of this self motion by visualising additional optical flow (Bruder et al. 2013). Then the movement behaviour that resulted from such visual stimuli was prompting passive resposnse in walking paths. For instance Bruder et al. (2013), superimposed the flow information to alter the perception of self motion that resulted in walking movements to be either faster or slower than it really is. Another work in the similar direction (Ishii et al. 2016), exploited the use of body worn HMD to influence movement by the manipulation of the visual content as shown in Figure 3.2a. In the work, a stereo sensor was first used to capture the real world scene while walking. Then by applying image processing, visuals of moving strips were added to the scene. The moving strip approach successfully induced *vection. Vection (Fisher 1930) refers to the compelling sensation of self-motion by a moving visual stimulus.* This for example can be experienced when waiting in a car at a traffic signal or when observing other vehicles nearby starting to move. Another visual variant that was applied in the same study investigated the effects of changing visual content. This variant then proved to influence walking paths based on where the user focused and fixated gaze while walking. Either of the two approaches (moving strips and changing content) in the work were successful in affecting both the path and direction of walking. Another work that also applied influences based on the principle of vection

visually augmented the ground planes of the walking scene (Furukawa et al. 2011). The study for such walking floor illuminated walking effects involved a pedestrian being led by a moving visual stimulus projected on the floor (Figure 3.2b) using an optical device. Thus using a projection based augmentation in a static setting, influences to both motion and direction of travel were achieved.

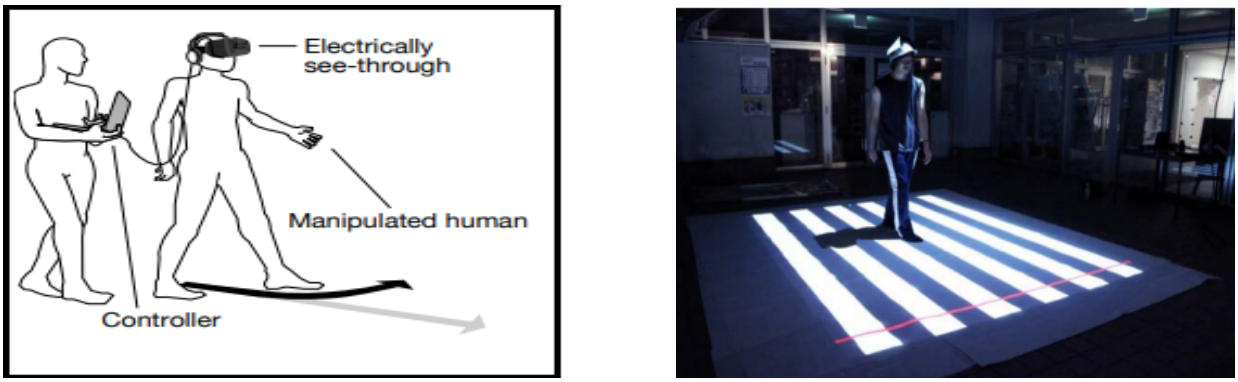


Figure 3.2: (a) The figure on the left shows how vection is applied to a HMD device (Ishii et al. 2016) while the right shows (b) the image of a person influenced in walking using light projection vection (Furukawa et al. 2011).

Much of the research that have focused on manipulating walking using a body worn HMD have emphasised on the ease of walking movement that is supported by the handsfree operation of headsets. Projection based approaches on the other hand are more prone to visual delays invariably occurring between cognition and the resulting moving action. While visual influences have been primarily studied with AR, non visual guidance methods have also been researched as an alternative guidance approach. Amongst them, vibro-tactile cues for navigation (Lindeman et al. 2004, Uchiyama et al. 2008) and motion guidance (Bark et al. 2014, Marquardt et al. 2018) have also been studied for their passive walking effects.

Redirected Walking in VR: Redirected walking (RDW) refers to a collection of motion manipulation techniques in Virtual Reality that makes it possible to move about in a physical space of smaller dimensions to achieve continuous walking in a very large virtual environment. Then if a person is walking in a virtual soccer field, it is possible that the virtual scene is actually slowly and imperceptibly rotated around the user when applying the RDW technique. This might cause him/her to walk in circles even though s/he thinks that walking is happening along a straight line path. Such visual manipulations are achieved by the control of user's walking path by applying transformations/shifts to the virtual scene (Suma et al. 2012). The maximum amount of shifts (gain) that can be applied for such manipulations have been established through empirical studies relying on psychophysical methods (Grechkin et al. 2016, Steinicke et al. 2009). The redirection techniques employed for RDW use an array of methods that include visual distractors (Peck et al. 2011), viewpoint manipulation (Bolte and Lappe 2015, Langbehn et al. 2016) or by using narrative events as opportunities to imperceptively manipulate user paths (Grechkin et al. 2015, Neth et al. 2012). As per the taxonomy proposed in (Nilsson et al. 2018), the common ways of redirecting users include translation, curvature, rotation and bending gains applied to the movement of the virtual camera as shown in Figure 3.3.

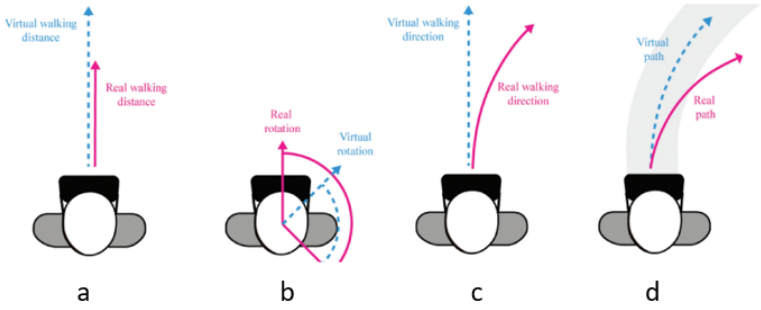


Figure 3.3: Four types of gains used for perspective manipulation (a) translation gain, (b) rotation gain, (c) curvature gain and (d) bending gain . The purple and the blue lines in the figure represent real and virtual transformations respectively (source: Nilsson et al. 2018).

Since RDW is a motion manipulation technique, the walking steps while navigating a virtual space for past time stamps is input to an underlying algorithm. Then, the approach approximates redirected walking paths by estimating the human steps made when using the VR headset. While there have been few approaches (Hirt et al. 2022) that have applied future motion predictions to manipulate RDW, such forecast based approaches have not received much attention to successfully influence motion in VR. This is largely due to the nature of human walks which is characterised by randomness, spontaneity and short term prediction complexity. Other RDW works that also tried to estimate the future walks approximated locomotion behaviour with data driven approaches (Strauss et al. 2020, Stein 2021).

Scene Aware Visualisations: While there have been fewer works, to the knowledge of the author, that have applied the results from scene awareness to influence walking; most works have embedded visual content based on scene motion to assist decision making in data analytics. For example, to augment sports videos in a post processed fashion, Chen in (Chen et al. 2021) applied data processing pipelines extracting postures of the player and position of ball. The output player and ball movement poses from the pipeline was then augmented to the video as player meta data in the visualisation step. A more advanced player movement visualisation and processing was achieved in iBall (Chen et al. 2023). For this a computer vision pipeline first detected the players and tracked them with a kalman filter. The iBall then automatically embedded visualisations highlighting players for a given sports scene. Furthermore, semantic segmentation was also applied in this work where the scene separated the background from its foreground to embed the visuals. While the above mentioned post processing works have augment visualisations only from a 2D perspective, incorporating context based visualisation into mixed reality is a lot more complex. This would require understanding the scene in real time as a prerequisite step to achieve this. Even when the term *Context Aware Mixed Reality* has been coined by Papadopoulos et al. (2021) and used often; such context applications require well researched pipelines where the scene data is first processed before placing virtual content.

Most mixed reality development platforms already incorporate basic scene understanding algorithms as a part of their standard toolkits. MR/AR toolkits like the MRTK [1], ARKit

[1] https://learn.microsoft.com/en-us/windows/mixed-reality

[2] and Google ARCore [3] render virtual content only after having detected the planes in the real world scene. The 3D geometry of the environment is perceived using the depth from the visual odometry SLAM algorithms present in these toolkits. Hence the detected planes improved the geometric context of the viewed AR scene. This has helped to achieve realistic collision effects for real objects (e.g., shouting objects in a walking game), occlusion rendering (i.e., blocking the virtual objects to place 3d scene models) and distance based light effects for virtual content in real scenes (Linowes and Babilinski 2017, Alfakhori et al. 2022, Kumaran et al. 2023).

Furthermore, some recent context based works have exploited AR capabilities beyond plane detection so as to include semantic and scene motion information. A framework to support context-aware interactions using scene specific data was proposed in (Chen et al. 2018). For this, a depth map from an RGBD sensor was fused with semantic information to create semantic 3D models. The improved 3D model was then used to improve both the real and virtual interactions in the game application. On the other hand, interpreting contextual data based on moving scene objects was tested in TransformerMR (Kari et al. 2021). In the work, Kari et al. (2021) processed the feed of a monocular RGB camera with a pipeline for scene perception, transformation and construction to create virtual experiences where real cars and pedestrians were replaced with virtual moving avatars. For this, the work performed a pose aware object substitution by first applying semantic segmentation and pose estimation onto vehicles in the scene. This was followed by the object impainting (Kim et al. 2019) step that removed cars and people from the scene. The last construction step then placed virtual avatars for the estimated 3D pose of the removed objects. For the mixed reality work that was demonstrated using a deep learning pipeline, the application successfully achieved real time performance at 15 fps.

Scene representation and its abstraction (as scene graphs) for AR content placement have been studied in recent virtual studies. MR context in retargetableAR (Tahara et al. 2020), was represented as a graph with scene objects as interlinking nodes. This graph was then used to determine how real objects would interact with virtual content added to the scene. In the graph creation step, semantic segmentation was applied to the 3D data of the scene. This was followed by the scene voxelisation step. Each object was then represented in the 3D space using oriented bounding boxes which was further condensed to a 3D scene graph. A recent work by Li et al. (2022a), extended the retargetableAR to support scene placement at run-time. This was demonstrated for story telling as a use case.

While not much works to the knowledge of the author has applied perception for safety in MR, safer methods to navigate in virtual reality was proved to be possible in (Cheng et al. 2019). The prototype in the work VRoamer, used a head-mounted RGBD camera to dynamically detect and perceive moving objects in front of a VR immersed user. Using an environment and motion pipeline, the scene and its geometry was initially identified. Any objects that stood out from the scene geometry was then mapped as an obstacle. This allowed users to safely move in the virtual world while avoiding collisions with objects present in the real world. Another work, the DreamWalker (Yang et al. 2019), applied perception along with VR authoring to create realistic walking experiences. For this an object detection system extracted moving pedestrians and objects in front of the headset using on the fly YOLO and depth sensing. The detected pedestrians were then substituted

[2] https://developer.apple.com/documentation/arkit
[3] https://developers.google.com/ar/develop

in VR using simulated content so that walking in the virtual world was a seamless experience as in reality.

3.1.3 External Influences

While walking there are always external factors and influences to motion that are in play due to movements happening in the near vicinity of the navigating persons. Simulation models that involve pedestrians movement in traffic spaces have often aimed to reproduce these influences occurring between persons and cars or between the different road users(Johora and Müller 2020,Ahmed et al. 2020). In such models, the walking behaviour of each person is modelled using a combination of social forces as proposed in (Helbing and Molnar 1995) or using the cellular automata model (Blue and Adler 1999). VR and Mixed reality have been used to both integrate such models (Kamalasanan et al. 2022b) and study influences of external walkers and other moving objects to once own walking. Most studies in this aspect has been on what factors influence the collision avoidance (Olivier et al. 2017 Berton et al. 2019) behaviours of a person when facing a crossing or a danger situation. Virtual environments (as in Figure 3.4) prove ideal in such works by supporting both the replication of real scene in virtual world and the recording of movement data to study walking behaviours.

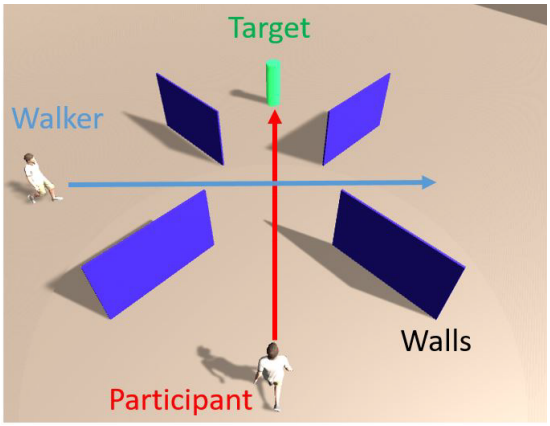


Figure 3.4: Walking motion interactions in VR study setting where a study participant crosses a walker in the presence of obstacles (walls) that occlude his/her vision (Berton et al. 2019).

Walking Studies and Virtual Environments: Motion studies (Krüger et al. 2024, Orschiedt et al. 2023) have been used to evaluate how walking people are influenced by the presence of both static and dynamic objects in a scene. Such studies have mainly observed what movement strategies would come into play when a person sees objects or other persons cross their path. Most studies have reported that a person would either stop, slow down or change his/her direction of walking when confronted with obstacles that would block his/her movement path. In the presence of static objects, obstacle avoidance for walkers was mainly explained by path adjustment as reported in (Huber et al. 2014). The person would move away from his/her current path seeing an obstacle in most scenarios tested in the study. This observation was further extended by Fajen and Warren (2003) to include moving obstacles to study collision avoidance. While most studies reported path and speed adjustments in their works, varying the speed has its own advantage in some cases. Speed adjustments often came with the benefit of maintaining the desired path so that spatial

re-planning for trajectories might not be needed when facing a crossing conflict. But when experimental constraints where put in place (for example by restricting view or crowding the walking spaces), braking was found to be required (Jansen et al. 2011, Moussaïd et al. 2011). Lack of sufficient walking spaces and a hurry to cross others could also make braking a necessity as noted by Cinelli and Patla (2008).

The bodily behaviour or spatial constraints applied to study collision avoidance are not the only factors that influence crossing decisions. The amount of attention given to any nearby crossing object measured using eye gaze also plays a role. For instance, Croft and Panchuk (2018) used the gaze information obtained using eye trackers while recording the resulting crossing behaviour for both constrained and unconstrained walking spaces. The study highlighted that making visual observations does play a significant role in the choice of a crossing strategy. In an another study, fixation was found to play a role in deciding who would cross first and also the probability of any resulting collision (Jovancevic-Misic and Hayhoe 2009) from the crossings.

The use of headsets with eye tracking feature and its ability to move handsfree while tracking each user position in 3D space has attracted VR and mixed reality as a good platform to study walking behaviours. The motion in virtual setting while wearing a headset has noted not to be significantly different from real world walking behaviour (Fink et al. 2007), hence proving the technology to be suitable for studying walking influences using AR. In an earlier work, Olivier et al. (2010) observed the crossing behaviours of persons with virtual pedestrians in a desktop based study. Using such an approach, the participants of the desktop study were able to both detect, estimate a collision and also anticipate whether they had to cross or give way. Similar avoidance behaviours have been studied in VR settings either when both interacting persons were represented virtually in the same VR environment (Podkosova and Kaufmann 2018b, Buck et al. 2019), or where persons interacted with a virtual 3D pedestrian avatar (Podkosova and Kaufmann 2018a, Nelson et al. 2023). Each walking interaction was then found to be influenced by the crossing situation, personality and gender (Knorr et al. 2016 Olivier et al. 2013) of persons in the study.

3.2 Discussion

A common element in each of the types of motion influences reviewed is that the methods are largely different based on the type of environment (real, virtual or mixed) and the visual medium and content used to influence walking. Each applied method might influence a walking person differently (e.g., redirected walking for virtual reality vs AR for navigation) and also the same technique might yield different results when applying it to a see-through AR headset like the Hololens. The degree of motion influence is expected to be different when using an active, passive or external stimuli intervention using an AR device.

Taking traffic context and driver movements for example, even when different AR visualisations and research works have focused on safety by indicating dangers in the scene (Schall Jr et al. 2013, Winkler et al. 2015), none of the works focused on how much would be the degree of influence to driving movements after having seen danger warnings. Much of traffic study literature has focused on evaluating whether seeing 3D visualisation was more effective than a 2D presentation or if people were actually paying attention to the 3D content shown on the displays (Tonnis et al. 2005, Schall Jr et al. 2013).

Also applying AR visualisation based on scene context has not received much attention. To improve safety based on traffic situation, a shared reality concept was introduced in (Peitso and Michael 2020). In it, tracked behaviour of real world objects in the vicinity of a vehicle was combined to show AR content. However the work lacked any technical details of how such a system would be feasible. Most AR studies for traffic safety only emphasized drivers and never considered pedestrians or influencing walking. Also none of the works to the knowledge of the author applied AR visualisation to passively influence the walking behaviour based on the safety scene context. Safety traffic scene context with reference to this thesis would mean the movement of other persons/objects (cyclist, vehicles) including pedestrians with conflicting motion walking nearby.

For an AR device like the Hololens to show a motion influencing visualisation to enhance safety, it should be able to first detect objects/persons walking nearby; track where these people would walk in a future point in time and show a scene context appropriate visualisation to make the Hololens user feel safe. Hence to identify the research gaps in achieving a safety based AR influence, we further subcategories our literature review to focus on key components- Context aware perception (*Localisation of persons/objects* and *Motion tracking*) and Augmentation (*Future visualisation*) of motion based on perception pipeline. The Figure 3.5 highlights the key review components researched within an AR context for this thesis. Also as both the mentioned components form an essential part of robotic motion perception research, to avoid exhaustive related works study; we only focus on reviewing works that are RGBD based or visualisation specific to identify gaps for this thesis.

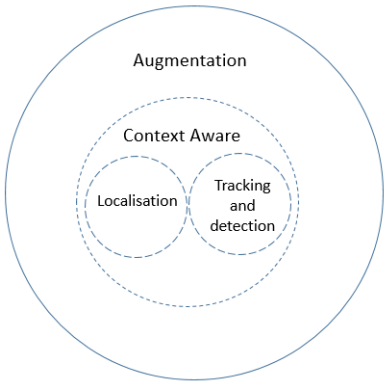


Figure 3.5: Scene Aware Mixed Reality for Motion Influence.

3.2.1 Localisation and Motion Tracking

Localisation of persons as seen from the perspective of an RGBD based sensor perception has been addressed using 3D object detection algorithms (Geiger et al. 2012) that have been developed in the field of robotics. In such approaches, given an RGBD image, the 3D position of the person, along with the dimensions of an enclosing 3D box is estimated. In pedestrian detection algorithms as in (Kollmitz et al. 2019), based on RGBD data, Kollmitz et al. extended a Faster R-CNN model (Ren et al. 2015) wherein 3D centroids of people were regressed using a deep learning network for an indoor mobility dataset. Another work (Linder et al. 2020), extended the Yolo v3 model by concatenating the depth along with RGB images and evaluated 3D centroid estimation of persons for synthetic data. Another RGBD version for YOLO was proposed in (Takahashi et al. 2020) applying extensions to

its backbone while regressing 3D centroids of people using 4 channel RGBD data directly. Most of the mentioned works focused on only estimating the 3D position of the persons and did not consider how the persons were oriented in the 3D space. Of the 3D detection approaches that accounted for the orientation estimates, Frustum Pointnets (Qi et al. 2018) provides estimates of not just the position, but also the orientation and enclosing 3D shape. While there have been more recent SOTA 3D detection approaches (Geiger et al. 2015) proposed recently, numerous works (Wang and Jia 2019, Wang et al. 2020) have focused on extending and improving 3D RGBD detection with F-Pointnet using the Lidar, camera and radars sensors. In a recent research, the JRD Dataset (Shenoi et al. 2020) was extended to both detect and track walking pedestrians in an indoor scene using the F-pointnet detection module.

Human motion tracking is also an essential component of motion perception where once a person is detected using localisation, s/he is tracked continuously to follow the detected person and his movements. Most modern trackers employ a tracking by detection approach to constantly track pedestrians in motion sequences (Greedy n.d., Roshan Zamir et al. 2012, Choi 2015, Dehghan et al. 2015, Yoon et al. 2016, Klinger et al. 2017). This method would come in two phases : (a) each person in the scene will be detection using pedestrian detection, and (b) detections from consecutive frames are associated to generate a set of movement trajectories. While detection can be tackled using the localisation block discussed in the previous section, tracking could be achieved when subsequent detection's of the same person or object are mapped and associated correctly. The tracking task could become a lot more complex depending on the number of persons appearing in the scene and other environmental factors. While previous works have dealt with optimisation methods (Berclaz et al. 2011, Dehghan et al. 2015) to handle such issues, not much has been investigated on how tracking performances for AR visualisation pipelines would differ. While some works have applied detection to AR headset sensors like the Hololens (Zeidler et al. 2023, Bahri et al. 2019), none of the works applied motion tracking for AR device data based on tracking by detection approach.

3.2.2 Visualisation of Future Paths

Once the motion of people in an AR scene are estimated, then visualising their current position or their future paths have been studied in automotive display research. An AR user interface that casts a virtual shadow of approaching pedestrians was prototyped and tested in (Kim et al. 2016). For the Head Up Display (HUD) based AR visualisation, the appearance of the shadow was varied to indicate both the point and direction of intrusion. Another recent study by Yu et al. (2023) focused on using human motion path visualisations to increase trust for occasional and unplanned robot encounters in public settings. For this, Yu et al. visually communicated the inferred motion paths of walking pedestrians from nearby robots onto the collocated HMD headset users. Future path visualisations have also been attempted in other domains to make early decisions based on visual data. An earlier work in predictive sports visualised AR future trajectory using a motion pipeline (Itoh et al. 2016). The pipeline implementation supported motion estimation in real time such that the future position of ball could be estimated well in advance.

3.3 Research Gaps

Based on the review of related works, the following research gaps were identified.

- **Perceiving pedestrian scene motion using AR device**

  There are only few works and methods that explored scene motion estimation using AR devices and its sensors. As each AR device would have RGBD sensors to create 3D maps of its surrounding scene to place virtual content; applying perception to the data from the device might be beneficial to overlay appropriate visualisations to influence walking.

  In this thesis, a motion perception pipeline that uses the RGBD sensors of an AR device (Hololens) is proposed. Using existing methods from robotic perception, the pipeline will detect and track persons walking nearby the headset.

- **Evaluating the motion inference capability of an AR headset**

  The accuracy of a pipeline that detects and tracks people is important if the objective is to apply appropriate AR visualisations based on nearby motion. So, if the motion pipeline is not able to accurately detect and track where others in front of a Hololens user are walking, the resulting visualisation applied based on inferred motion will have inaccuracy in creating motion influences.

  The motion trajectories from a tracking by detection pipeline gives a quantitative and qualitative estimate of how good the estimated scene movement is. Hence by using the tracked trajectories of detected people of the Hololens, a method of quantifying the accuracy of estimated walked paths for the AR device is explored. This has not yet been done in any previous studies.

- **Ego influences of visualising future motion based on pedestrian scene movement**

  A future guess of which path persons takes predominates navigation decisions of others walking nearby. Especially in the event of potential collision, this estimate s/he makes controls the way how the conflict spot would be avoided. Showing the future path of others in AR as extra information could either prompt one to reinforce or contradict his/her conflict avoidance strategy.

  In this thesis, a novel method of studying the path choice influences when visualising the future path during a motion conflicts is proposed. The approach that focuses on walking influence can be studied from preferred motion trajectories that people might decide to take while avoiding collisions with AR futures.

- **Studying other influences that can be captured with AR Sources**

  When walking, people prefer to make movement decisions not solely based on nearby persons, but also from the presence of other traffic participants such as cyclists moving nearby and/or from environmental factors. The presence of a virtual traffic light indicated to them could be once such example of an environmental component.

  In this thesis, a novel method of studying such external influences is proposed by either applying an active influence using AR (e.g., virtual traffic light) while on foot or by measuring the impact of external influences (e.g., presence of nearby cyclist). Both the influences are further studied using the trajectories of the involved agents.

4 Dataset and Evaluation Methods

In this chapter, we survey and discuss the datasets and evaluation metrics that have been applied in the scope of this thesis to design, evaluate and validate the methods of AR influences. To train and evaluate pedestrian detection, the simulated shared space dataset is used. The tracking performance of the motion perception pipeline is evaluated using the IKG pedestrian tracking dataset. The detection, tracking and conflict based metric are further used to evaluate the motion pipeline and AR influence methods introduced in this thesis.

4.1 Simulated Shared Space Dataset

To create an RGBD dataset using the Hololens sensors, the Simulated shared space (SSS) dataset was created using the device in the research mode (Ungureanu et al. 2020). In the experimental arena designed to capture the data, the Hololens overlooked an open space indoor scene that contained chairs and benches (Figure 4.1). The objective of the data capture was to create a pedestrian RGBD dataset as seen by a ego user for an indoor space. The motion and interactions in the space was expected to mimic the activity and behaviors that might take place in an outdoor shared space. For this, the arena included floor markings that were synonymous to car lanes and street furniture (tables and chairs) to recreate social interactions. Also the indoor space dimensions roughly matched the Hololens depth sensor range ($\approx$ 7 meters) to capture quality RGB and depth images for the walking scene.

In the data collection campaign, three participants along with three volunteers completed an enacted walking sequence with social interactions. At the start of the experiment, each participant was shown a picture of the shared space and instructed to use the marking on the floor to assume the existence of passing cars in the environment. The narrative that was used to guide the participants imitated the walking journey of persons shopping in a German farmers market. All persons who participated were instructed to act in the most natural manner pretending to navigate in a shared space while meeting friends. The volunteers helped the participant to move in the space and either walked with them or intentionally created conflicted walking movements.

An external static camera was setup overlooking the ego Hololens to capture the different recording sessions and to support the documentation of the dataset. Each data compare session lasted for three minutes with the Hololens capturing depth and Image using the Research mode util application[1]. Following each data capture session, the data was manually downloaded to store the RGB image, depth and camera pose. The Figure 4.2 illustrates an example RGB and corresponding depth captured by the ego Hololens overlooking the walking scene with participants and volunteers.

[1] https://github.com/microsoft/HoloLens2ForCV

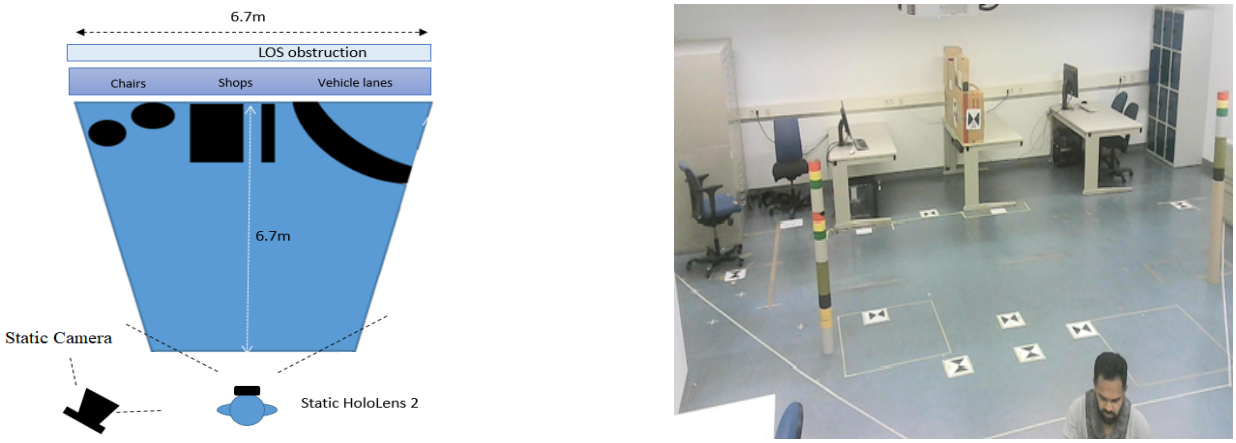


Figure 4.1: The figure on the left shows the data capture plan where an ego Hololens user looks at a scene with chairs, retail benches and lanes drawn on the floor. The figure on the right depicts the indoor scene implementing the plan of a shared space.

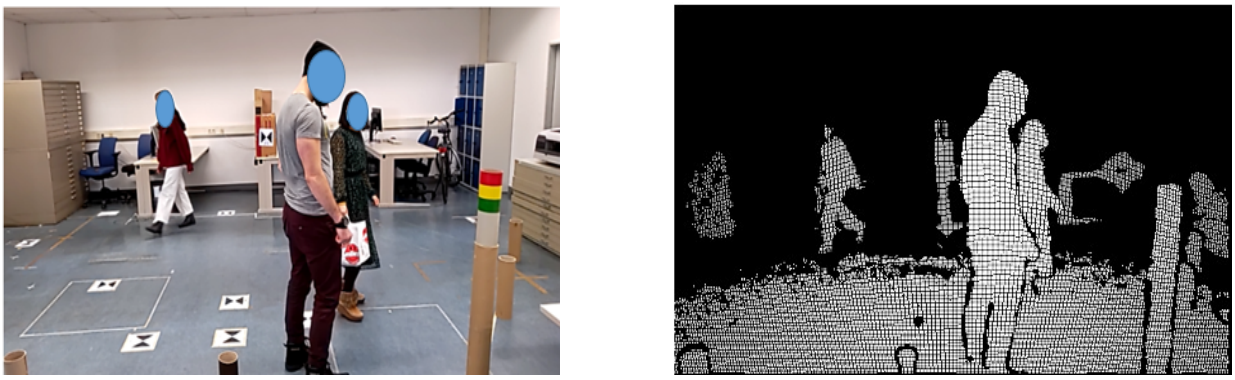


Figure 4.2: Images of the dataset captured using the Hololens RGB camera(left) and Depth camera(right).

A large dataset of image and depth streams with pedestrian movement was created from the data collection campaign. The following pedestrian motion characteristics (Table 4.1) were recorded in the dataset:

Characteristics	Count
Pedestrian Collisions	7
Grouping	9
Total Duration (min)	18

Table 4.1: Pedestrian interaction characteristics of the SSS Dataset.

Semi Automated Pedestrian Annotation: A semi-automated labelling method has been developed in this thesis to annotate position of pedestrians in 3D captured in SSS Dataset. The method was used to created bounding boxes of persons in the Hololens 3D scene. The annotated data was then further used to retrain 3D pedestrian detection algorithms used in this thesis.

The Figure 4.3 details the annotation pipeline used to created bounding boxes by fusing data from both the 2D images and 3D pointclouds. For each RGBD image captured using the Hololens, the 3D scene and the RGB images are processed simultaneously in the pipeline. The 3D point cloud of the scene is first removed of any ground planes by using the RANSAC algorithm (Derpanis 2010). Simultaneously pedestrians in the scene are detected using the YoloV3 (Redmon and Farhadi 2018) image detector. Hence by using the ground plane filtered pointcloud and yolo results, all points that could belong to a pedestrian are identified by separating pedestrian points from non pedestrian points. Also as the data capture arena for the SSS Dataset also contained tables and benches, a region growing approach was used to segment out persons in 3D from any artifact noise. For every segmented pedestrians, a bounding box was estimated that enclosed the 3D points of the person in the scene. The final step of the semi-automatic annotation included correcting any errors in estimated boxes using manual correction. In this step, each of the 3D pedestrian bounding boxes estimated following the region growing segmentation was manually corrected for orientation errors using Labelcloud (Sager et al. 2021).

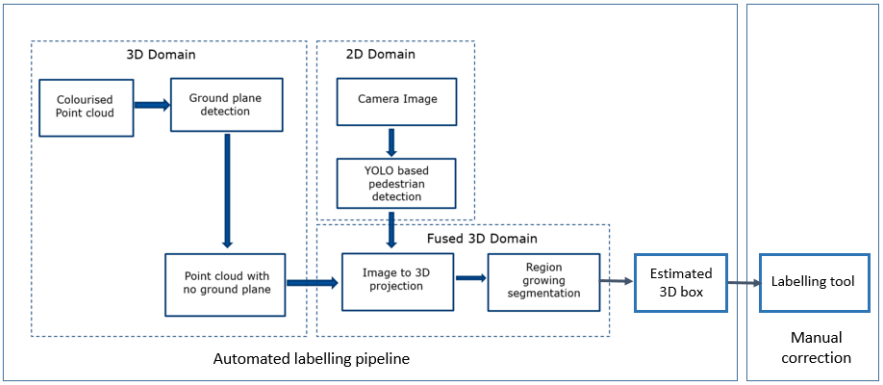


Figure 4.3: Semi automated pipeline for labelling pedestrians for SSS Dataset.

4.2 IKG Pedestrian Tracking Dataset

The IKG tracking dataset consists of motion trajectories of pedestrians walking in an indoor lab recorded at Appelstraße 9a, Hannover. The indoor space (Figure 4.5) used for the data capture was characterised by a free walking environment of dimensions 6m x 5m and was largely open. The objective of the data collection campaign was to create an ego pedestrian RGBD tracking dataset using the Hololens 2.

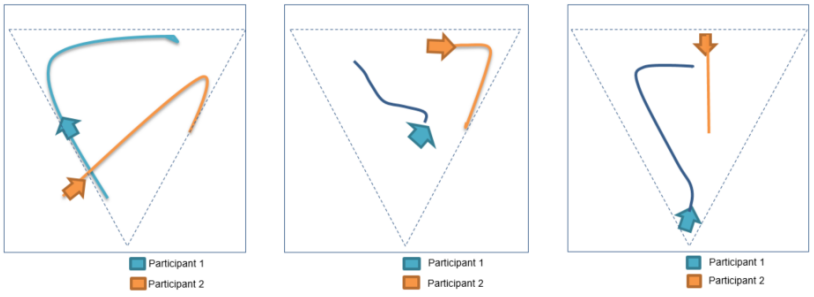


Figure 4.4: Walking sequences captured as part of the tracking dataset.

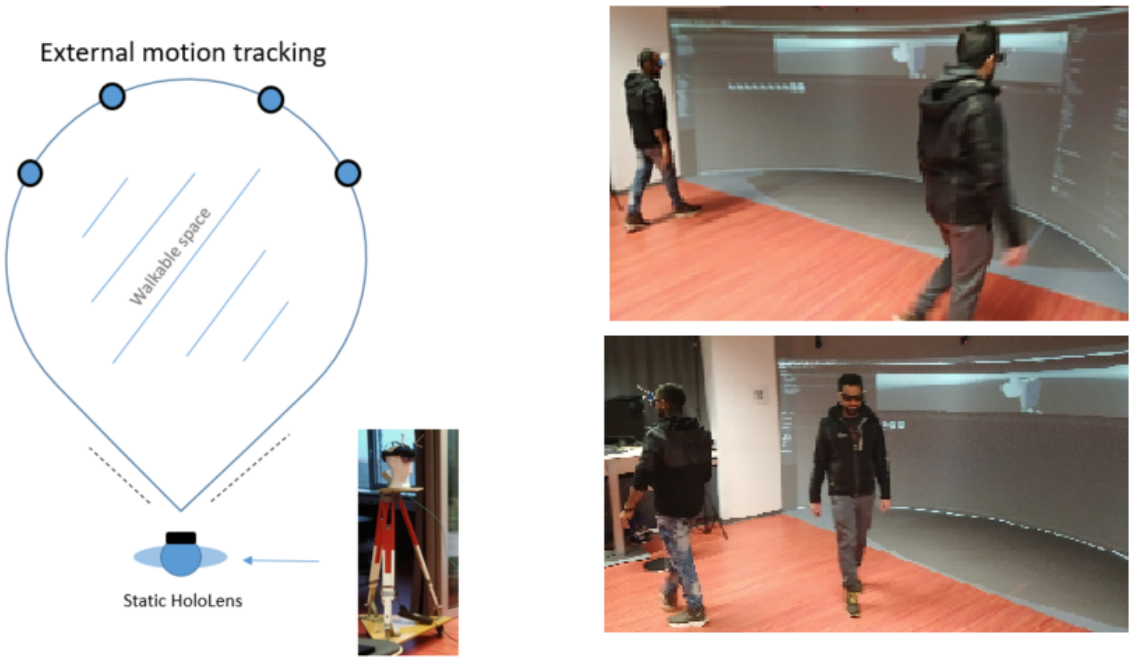


Figure 4.5: The Indoor tracking scene capturing walking pedestrians using both optical motion tracking and the hololens.

For this, the Hololens was placed on a mannequin and positioned at a height of 1.5 m above the ground viewing the open walking space. Two different volunteers were invited to the tracking data campaign and asked to walk in specific walking sequences as illustrated in Figure 4.4. Each of the volunteers worn eye gears that embedded optical markers on them. The markers were used to capture ground-truth motion and their tracks in the scene. During the capture, the motion in the lab was recorded using two sources- Hololens and optical motion tracking. While the research mode sensors (RGB and Depth) of the Hololens captured the movement of the volunteers in front of the device, their motion was also captured simultaneously using a high precision optical tracking system[2] to create groundtruth data.

The resulting captured dataset consisted of motion tracks from two walking pedestrians in the scene moving in different crossing configurations (path crossing, side-by side motion or 90 degree conflict). Using post-processing the Hololens RGBD data was time synchronised to the optical tracking movement data as explained further in section 6.3. Each timestamp of the captured RGBD data from the Hololens then recorded a groundtruth track id, a 2D position, an RGB image and its corresponding depth pointcloud.

4.3 Evaluation Metrics

To evaluate the methods proposed in this work, different metrics are detailed based on the component examined. As detection of persons in 3D, tracking their movement and evaluating the safety of walking with AR visualisations are the primary research focus of the work; different metrics from robotic perception and traffic safety are used in this work.

[2]https://ar-tracking.com/en

To perceive pedestrian movement accurately, a motion perception algorithm should report on how many persons were detected correctly and how may false positives (false alarms) were produced. With respect to tracking them, it should support one-to-one matches, one-to-many matches, many to one matches, and should scale up to larger test areas without loss of tracking capabilities. And finally to quantify safety while walking when viewing AR visualisations, the chosen metric should estimate if collision severities have increased or decreased with AR content.

4.3.1 Detection based Metric

Given the predictions from an object detection algorithm, the output object class could either be (Fig. 4.6): True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Here "positive" means prediction of being object of issued classification, while "negative" means prediction of not being object of interest. "True" and "False" make a judgement, whether each prediction match ground truth or not, which is decided from either their IoU or threshold of prediction based on the ground truth data. For a 3D object detection approach, this would mostly be the 3D IoU of the predicted boxes.

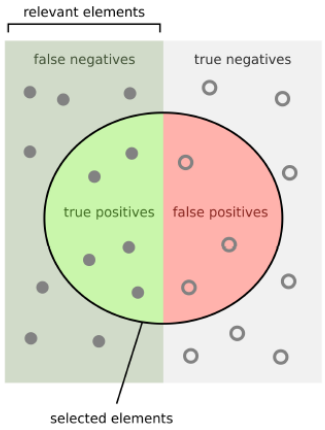


Figure 4.6: TP: items correctly labelled as belong to the positive class. FP: items mislabelled as belong to positive class. TN: items correctly labelled as not belong to positive class. FN: items mislabelled as not belong to positive class.

Accuracy, Precision, Recall and Average Precision : These four metrics are used to evaluate how good an object detection algorithm is based on the total number of positives and negatives predicted.

The accuracy (Equation 4.1) is evaluated by computing the ratio of total categories samples with the sum of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

However, in object detection evaluations, the accuracy is not be taken into account due to uncertainty of TN. Instead, two other metrics - precision and recall are also used as object detection metrics.

The precision for a class is the number of TP divided by total number of elements labelled as positive:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

The recall is defined as the number of TP divided by total number of elements actually belong to positive class:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Obviously, an optimal object detection model should gain higher score of precision and recall, but there is an inverse relationship between them. It's possible to increase one at the cost of reducing another one. Therefore, average precision (AP) is usually used as the evaluation standard to combine precision and recall.

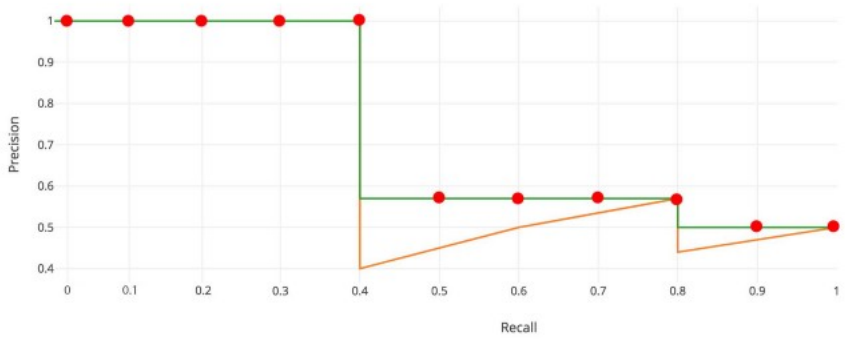


Figure 4.7: Recall value is divided from 0 to 1.0 into 11 points. The interpolated precision is marked as red dot, which takes maximum precision over recalls > r.

The Pascal VOC 2007 challenge (Everingham et al. 2010) provides a standard to compute AP (Figure 4.7). It sets the IoU threshold = 0.5, and calculates every precision over 11 recall levels $\mathbb{A}$=[0 : 0.1 : 1], the interpolated precision $P_{interp}(r)$ that takes maximum precision over all recalls > r is denoted as:

$$P_{interp}(r) = max_{\mathbb{A} \geqslant r} P(\mathbb{A}) \tag{4.4}$$

The average precision is computed by averaging all interpolated precisions:

$$AP = \frac{1}{11} \sum_{r \in 0, 0.1, \ldots 1} P_{interp}(r) \tag{4.5}$$

4.3.2 Tracking based Metric

A tracker that follows multiple objects in a scene should at all points in time find the correct number of objects present and estimate the position as precisely as possible. Each object should be assigned a unique trackID that would stay consistent throughout a sequence (even in the presence of occlusion e.t.c.,). Hence the evaluation procedure for the sequence (t_1

....t_n) should correctly estimate the *tracking precision* expressing how well exact positions of people are estimated and *tracking accuracy* which would indicate on how many mistakes the tracker made in terms of misses, false positives, mismatches and failures to recover tracks.

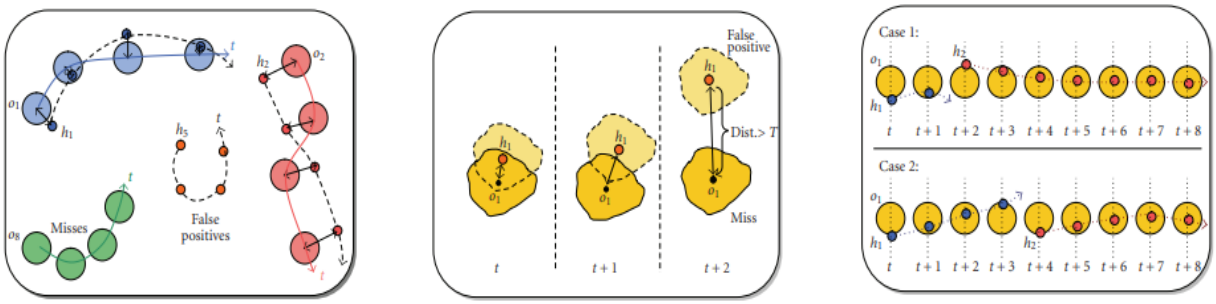


Figure 4.8: The figure on the *left* illustrates tracker hypothesis to object mapping as illustrated in (Bernardin and Stiefelhagen 2008), the figure in the center depicts correspondence discontinuity as distance exceeds threshold T. The figure on the right depicts a scenario on how mismatch count (Case 1-one, Case 2-one) calculated for correct mapping between h_2 and o_1 remain the same considering the length of the switched segments.

To establish a continuous mapping between object hypotheses from an object detector (h_1,.....,h_m) and the real objects (o_1,.....,o_m), only valid correspondences should be considered while tracking and the resulting tracks should be consistent over time (Figure 4.8). For valid correspondences, a certain threshold T determines the conceptual boundary beyond which the tracker misses the object. In the case of the 3D detection, the bounding box centroid distances between the object and hypothesis should be above that average width of a person for valid detection's to be tracked. Hence the value for this threshold T cannot be generalised and does depend on the distance measure used, the application task and size of the objects tracked in the scene.

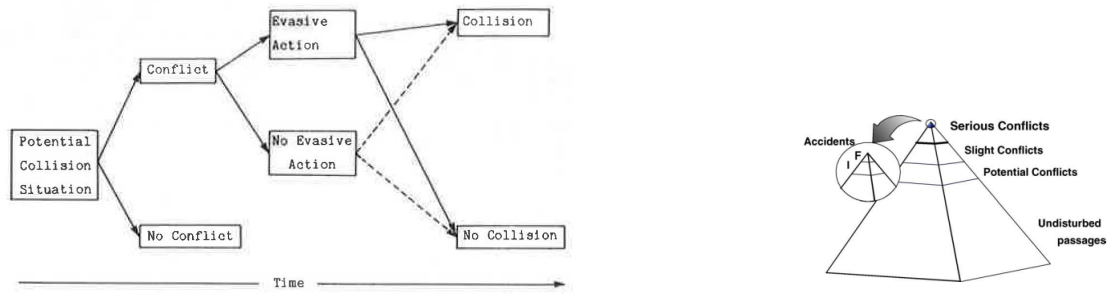
To keep track of the mismatches, a list of object-hypothesis mapping is maintained during tracking. For a set of mappings $M_t = \{(o_i, h_j)\}$ made upto time t, a new correspondence $\{(o_i, h_k)\}$at *t+1* that contradicts, would update the mapping set to $M_t + 1 = \{(o_i, h_k)\}$ increasing the mismatch count in the process. With this, the tracker keeps track on the number of mismatches while also maintaining the updated mapping relationships.

The performance of object tracking in this thesis is evaluated using the CLEAR MOT metrics (Bernardin and Stiefelhagen 2008). This metric consisting of multi object tracking accuracy (MOTA) and multi object tracking precision (MOTP). The MOTA represents the number of TPs, FPs and the Id switches (IDs) over n frames computed as :

$$MOTA = 1 - \frac{\sum_n (FP + FN + IDs)}{\sum_n GT} \tag{4.6}$$

MOTP describes how well a tracker can localise TPs which are computed in the 3D object space. The MOTP expresses the percentage of well-localised TPs, i.e., detection's having 3D distance to their corresponding GTs smaller than a threshold $\epsilon_{3D-MOTP}$

$$MOTP_{3D} = \frac{\sum_n I(dist(S, S_{ref}), \epsilon_{3D-MOTP})}{\sum_n TP} \tag{4.7}$$



(a) (b)

Figure 4.9: (a) Conflicts and collision generation (Allen et al. 1978), (b) Pyramid of interactions (Hydén 1987) for road users.

where I is indicator function with value of 1 if $dist \geq \epsilon_{3D-MOTP}$ and otherwise 0.

4.3.3 Conflict based Metric

A traffic conflict (Svensson and Hydén 2006) is a *situation where two or more road users approaching each other in space and time to an extend that collision would be imminent if there is no change in movements.* Conflict becomes an essential part of the collision avoidance (Figure 4.9a) that usually would arise when the involved agents sense a potential collision and then try avoiding it. This could be done by resorting to evasive action like moving away or changing onces motion speed. However, in the absence of any evasive action, each conflict would result in a collision.

Studying conflicts and its underlying reasons would helps to explain the severity of the traffic interaction amongst road users using the safety pyramid as proposed in Hydén 1987. As per the pyramid definition, the most severe conflict would be synonymous to collision while lesser severe encounters would mean lesser difficulty to cross each other. Also recent works that have evaluated safety for shared spaces have estimated threats based on the number of traffic conflicts due to pedestrian interactions (Orsini et al. 2023, Chen et al. 2017).

Surrogate safety indicators have been used in traffic safety studies (Johnsson et al. 2018) to evaluate the severity of interactions when two traffic agents come close to each other. Indicators like Time to collision (TTC) and Post Encroachment time (PET) estimate a temporal distance between the two agents when a collision becomes evident. These metrics then quantify the (a) the nearness of road users, and (b) potential evasive actions that the road users took in the event of a conflict. Both the metrics are further explained in detail.

Time to Collision (TTC): The TTC concept, which was introduced in 1971 by the US researcher Hayward, builds up on a constant motion speed and direction assumption for predicting the temporal distance to a potential collision of two road users in the future. In (Miller and Huang 2002), a procedure to calculate the time to collision between to traffic agents is detailed. For this, the data considered the initial positions of the agents along with its speed and direction.

Given the initial points as in Figure 4.10, the intersection points of the two agents are estimated by the equations:

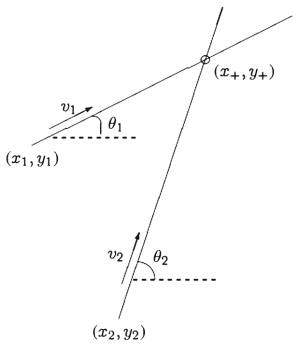


Figure 4.10: Simplified calculation of the intersection points given initial position of agents at (x_1, y_1) and (x_2, y_2).

$$x_+ = \frac{(y_2 - y_1) - (x_2 * tan\theta_2 - x_1 * tan\theta_1)}{tan\theta_1 - tan\theta_2} \tag{4.8}$$

$$y_+ = \frac{(x_2 - x_1) - (y_2 * cot\theta_2 - y_1 * cot\theta_1)}{cot\theta_1 - cot\theta_2} \tag{4.9}$$

Once the intersection point is estimated using Eq 4.8 and 4.9, the collision time is computed based on the current speeds of agents.

$$TTC = \frac{d}{v_2 - v_1} \tag{4.10}$$

where d is distance between them when either agents have reached the computed intersection point and V_2 and V_1 correspond to the speed of the involved agents.

Post-encroachment time (PET): As introduced by Allen et al. (1978), PET measures the actual temporal distance when the trajectories of two road users cross. In other words, PET is defined as "the minimal delay of the first road user which, if applied, will result in a collision course and a collision" (Laureshyn et al. 2010). Low values of PET reflects severe traffic conflicts and PET=0 represents a crash.

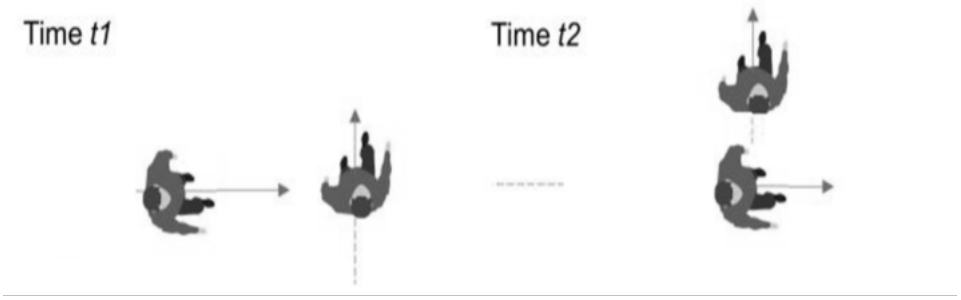


Figure 4.11: The time instance t_1 when the first person enters the conflict zone and t_2 when the first person leaves the conflict zone and the second person enters it.

The value for the encroachment time can be calculated based on based on the time it takes the agents to pass the crossing point as in Figure 4.11. The value for this metric is computed as:

$$PET = t_2 - t_1 \tag{4.11}$$

Due to the unpredictable walking behaviour of humans, it might be hard to quantify the safety interaction behaviours in this thesis solely based on the above discussed metrics, hence motion dynamics (speed and distance) of the involved agents are also studied for the work.

5 Influence Framework & Scene Perception

The section details on both the general framework and its perception component of how AR can influence the walking behaviours. For this in the first section we first explain our influence framework and its pipeline. In the subsequent sections we explain the components that were identified for the framework briefly. While this chapter mainly focuses on the primary component of the pipeline - pedestrian perception, the future path visualisation of the scene that is also part of the pipeline is detailed in the subsequent chapters.

5.1 Influence Framework and General Pipeline

Aiming at influencing the walking behaviour using AR, our approach developed is focused to understand how seeing the future walking paths (in AR) of other pedestrians walking in front of a Hololens would affect the path choices of the headset user. For this both the RGBD data and the visual interface of the AR device is used in the method.

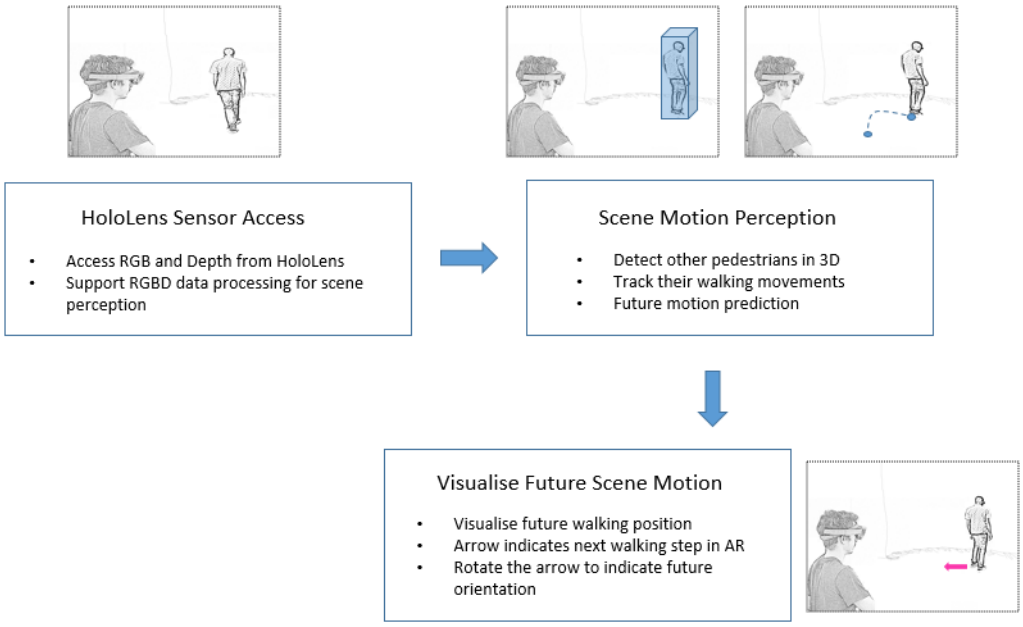


Figure 5.1: The overview framework for the proposed AR motion influence.

As indicated in Figure 5.1, initially both the RGB and depth data from the device is accessed simultaneously as a user looks at a scene. The data is then processed to understand pedestrian movement happening in the scene. For this scene motion perception is applied where different pedestrians walking in front of the device are localised. Once localised using data from both the past and current detections, the pedestrian trajectories of neighbours is approximated using tracking. Lastly, to create visual influences based on scene motion, the tracked paths can be first predicted and then visualised in AR. The future path visualisation then shows the next walking steps of others in a Hololens scene. The future steps

of neighbours are augmented using a 3D arrow such that the orientation indicates if the person would turn or continue to walk straight.

To explain the complete pipeline at a component level, the Figure 5.2 indicates the sensor access, scene perception and visual communication modules as highlighted in green & grey and red. Amongst the two perception blocks- motion interpretation and its prediction, the interpretation block has been given a higher priority in this thesis. The prediction block is intended to demonstrate how motion of others in the scene can be predicted with high accuracy given that the pipeline detects and tracks walking. Hence in our work we make a strong assumption that motion prediction algorithms like (Cheng et al. 2023) could be easily integrated into the influence pipeline with minor impacts to the overall proposed influence method.

The pipeline for the work follows a serial architecture where data from the Hololens is first accessed (*sensor access*) and then passed on to interpret the movement of pedestrians walking in the scene (*perception*). Having interpreted the future using the data; the *communication* component then overlays the future motion for the scene using the AR visual display to influence the walking choices of the Hololens user. Hence our work can be best partitioned into sections for innovations in each of the modules. While this chapter explains how scene interpretation is achieved using the Hololens, its visual communication and evaluation are detailed in the following chapters.

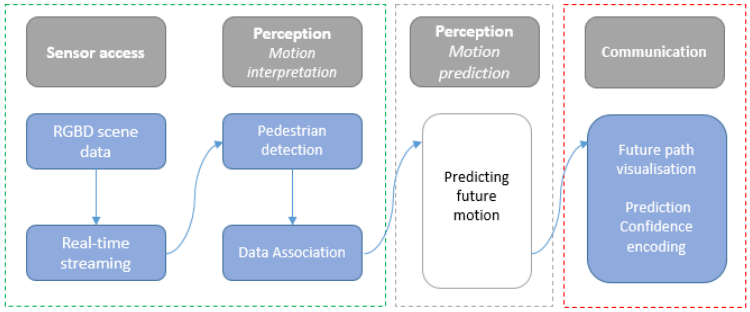


Figure 5.2: The figure highlights the components (grey blocks) and the flow of information (blue blocks) for the motion influence pipeline with sensors access and perception modules in green & grey and the visual communication module in red.

5.2 Sensor Streaming and Hololens

The Hololens 2 is a mixed reality device that supports running AR applications written using the Universal Windows Platform (UWP) [1]- programming interface for Windows client applications. The Device Portal [2] that is supported by Hololens is a web based tool that supports connecting to the device and streaming both audio and the video feeds while in operation. However the portal does not provide real-time data access to the Hololens depth sensor-Long Thraw Depth. Also the research mode utility application that is published along with the research API (Ungureanu et al. 2020) only supports recording the sensor data as files on the headset. This makes it difficult to run motion perception algorithms on the device in real-time while worn by the user. Hence to fill the gap and to provide access

[1] https://learn.microsoft.com/en-us/windows/uwp/
[2] https://learn.microsoft.com/en-us/windows/uwp/debug-test-perf/device-portal

to the sensors, a TCP connection based sensor streaming component was developed. This allowed streaming RGB and Depth images in real-time using the ROS (Robot Operating System) platform.

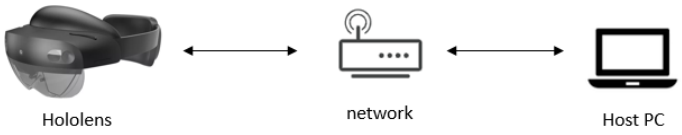


Figure 5.3: Hololens running UWP application connected to streaming Host PC over the network.

The streaming component (Figure 5.3) consists of a mixed reality application running on the Hololens and a receiver client & its server ROS instance on a host PC. The receiver on the host machine then subscribed to the RGBD data send by the UWP application and passing it onto ROS (as topics [3]) for further perception operations. The Hololens application is programmed using C++ and the receiver client instance is written in python. The noetic [4] distribution of the operating system ROS is used in the current work. All the motion perception components that are detailed in the subsequent sections are then run as independent modules subscribing to the hololens to access the raw data.

Hololens Coordinate System: To detect persons and track motion, we use the 3D scene point cloud and images from the AR device. The scene depth captured by the device has a limited range (≈ 5 meters) both observed manually and as reported in (Hübner et al. 2020). The small field of view of the sensor is however not considered a limitation in our current work. Also the 3D point cloud data from the device follows a sensor coordinate system as shown in Figure 5.4. All walking motion of persons in front of the device is then captured in the XZ local coordinate system for this thesis.

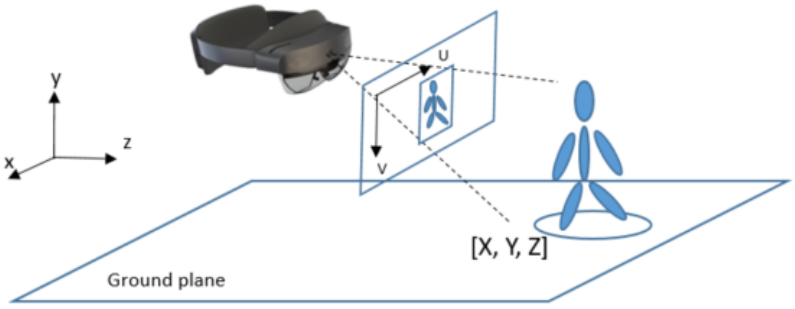


Figure 5.4: The Hololens coordinate system, with the Z axis pointing in the viewing direction.

5.3 Perception- Pedestrian Detection

Using the RGBD data from the Hololens device, two pedestrian detection algorithms **Frustum Pointnet** and **MaskRCNN** are applied in our work to detect pedestrians walking

[3] https://wiki.ros.org/Topics
[4] https://wiki.ros.org/noetic

in front of the AR headset. The choices two algorithms for the Hololens is to compare and evaluate how well RGBD data can detect moving pedestrians using both methods. Following detection, both F-Pointnet and MaskRCNN are compared for pedestrian tracking performance. For this the 3D detection data is then passed onto the tracker to associate the detections in each of the subsequent timestamps as pedestrian tracklets. In the following sections we explain both the detection approaches in detail.

5.3.1 Frustum Pointnet Detection

The pioneers in detecting pedestrians, cyclists and vehicles using a subset of input point cloud data as known as frustums is the Frustum Pointnet (F-Pointnet) (Qi et al. 2018). The network uses a high performing image detector (e.g.,YOLO) to first detect and identify people in the input images captured from the scene. The output from the 2D image detector is then used to localise the position of detected people using its corresponding 3D point clouds. The different stages that are part of this 3D detection approach as illustrated in Figure 5.5 are as follows:

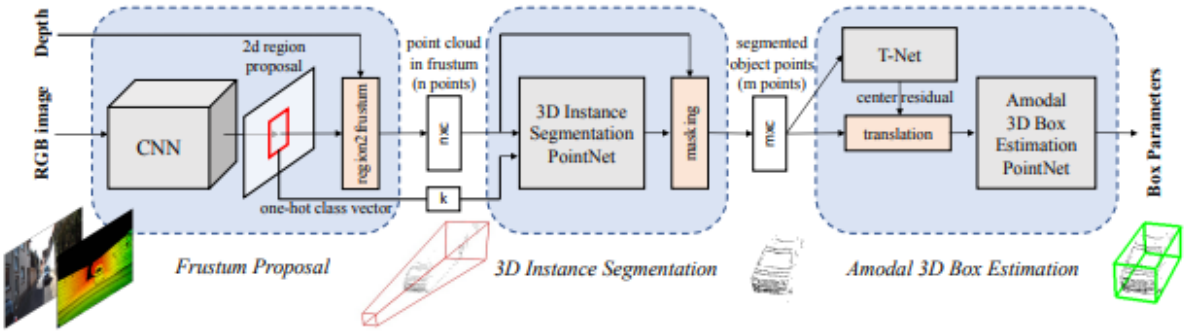


Figure 5.5: Overview of the Frustum Pointnet illustrating the main steps - *Frustum Proposal*, *3D instance segmentation* and *Amodal 3D Box estimation* (Photo: (Qi et al. 2018)).

Frustum Proposal Given the RGBD data of the scene from the Hololens, this step first passes the 2D images to a person detector. The output of the image detector (as 2D bounding boxes) along with the point cloud data is used to filter out all those 3D points that might possibly contain a detected person. For this, each of the 3D point of the input scene $P = \{p_1, ..., p_N\}$ is projected to the RGB image plane to obtain the pixel representation $p_i^{img} = (u_i, v_i)$ of the scene. Using the information from 2D image bounding box, all the points that lie inside of the bounding box are identified using Equation 5.1.

$$\begin{aligned} u_{min} \leq u_i \leq u_{max} \\ v_{min} \leq v_i \leq v_{max} \end{aligned} \tag{5.1}$$

where u = $\{u_{min}, u_{max}\}$ and v = $\{v_{min}, v_{max}\}$ represent the bounding box corners of detected persons in the image.

This step then returned a subset (frustum points) of the pointcloud with 3D information of all points that could potentially contain a person. The extracted frustum points $P' =$

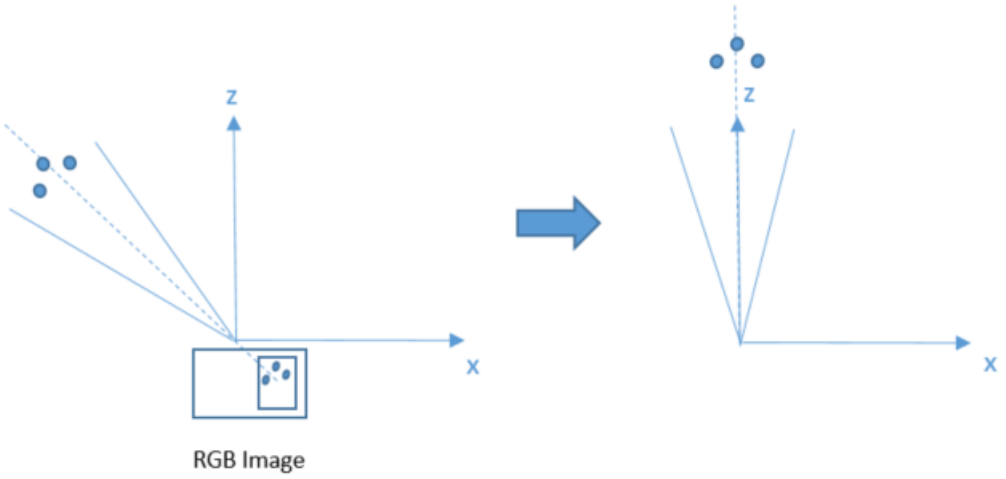


Figure 5.6: The figure (left) illustrates a detected bounding box on the image and frustum points that fall inside it. Frustum point cloud rotation normalises the frustum points using frustum angle. The rotated angle is then represented with an axis orthogonal to image plane.

$\{p_1, ...p_{N'}\}$ are then randomly sampled to obtain a uniform point representation for neural network based operations that are used in the later stages. The sampling step is followed by a normalisation operation where the subset point cloud is rotated around an angle- the frustum angle. The frustum angle represents the degree of tilt that the frustum pointcloud makes w.r.t the camera center. As illustrated in Figure 5.6, based on whether the person is detected to the right or far edge of the image center, pointcloud points that fall inside the detected box could be oriented differently in 3D with reference to the camera vertical axis.

The frustum rotation step attempts to achieve rotation invariance in detecting people in 3D. Hence, irrespective of the position of the person on the camera image and how the extracted frustums might orient following frustum rotation, all 3D detections happen in an orthogonal plane along the camera center. To compute this rotation angle, the bounding box center of the person in the image plane is computed. Then by reverse projecting the center, the angular orientation of the frustum point cloud w.r.t to the vertical axis is obtained. The rotated frustum cloud is further passed to the subsequent stages to localise the position of persons using a network based segmentation and box estimation method.

Segmentation and 3D box Estimation In this step, the pointcloud processed in the previous step is passed through a combination of smaller networks to first identify 3D points belonging to a pedestrian and then estimate their position and orientation respectively. The instance segmentation sub-network inputs the view point normalised pointclouds and identifies points belonging to pedestrians. This differentiates pedestrian points from non person 3D points and removes background noise. The network then returns the semantically classified points along with the probability scores when it identifies people in the scene. The instance segmentation network used in the F-Pointnet is based on the Pointnet (Qi et al. 2017) architecture. The Pointnet network that forms the backbone of many state of the art object detectors (Lang et al. 2019, Shi et al. 2022) applies multi-layer perception (MLP) to interpret both the global and local features in an input cloud. These features are then used in inferring the position and physical attributes of the identified persons by F-Pointnet.

In the last stage of this 3D pedestrian detection algorithm, all the points classified to the object class *person* are further regressed to a oriented 3D bounding box to fit to the segmented points. Before the final regression stage, the points are first transformed by the estimated centroid (x,y,z) obtained from the global pointnet features. With this transform, all points are now expressed in a local cordinate system. These locally represented points are then used to approximate the size of detected persons. For this a simplified version of the T-network outputs residues relative to the estimated centroid and pedestrian dimensions relative to predefined pedestrian anchor boxes. The final output of the network is then the estimated 3D position, the box dimensions (width, height and length) and the orientation (θ) of persons standing near the Hololens.

5.3.2 MaskRCNN RGBD Detection

In the second algorithm that is used in our work to detect pedestrians, we use the popular image segmentation network- MaskRCNN (He et al. 2017) to first segment pedestrians in RGB and then use the pointcloud information from its corresponding depth to localise segmented pedestrians in 3D. Unlike the previous F-Pointnet that performs 3D detection as a two step method to first reduce the search of 3D point cloud and then detect persons, this method uses the data from the Hololens sequentially. The primary MaskRCNN network is responsible to detect the person, and the extrinsic registration is only used to lift the 2D pedestrian pixels to 3D space. In the absence of an accurate registration, given a segmentation mask of the detected person from the image; the approach could fail to localise the persons accurately in 3D. This could primarily occur when a few nearby 3D points would wrongly overlap with the mask, resulting in erroneous 3D positions of the persons in the scene. As an accurate registration between the RGB image and depth is an important prerequisite to avoid such errors, we first accurately calibrate the RGBD sensors of the Hololens in our work.

RGBD Calibration The Hololens 2 is a commercially available mixed reality device targeted to support research and business use cases. Hence the depth sensor of this ToF device (Hololens 2) required re-calibration as its sensors were factory calibrated and contains errors (as in Figure 5.7). Most calibration techniques used to calibrate RGB and Depth cameras use the homography (Dubrofsky 2009) based camera calibration technique available with the Matlab calibration toolbox (Bouguet 2004). Such calibration approaches correct the errors in depth using the measurement of known landmarks (e.g., chessboard) and 2D-3D correspondences. This approach was however found to be unsuitable for the Hololens as the low resolution of the captured RGB and depth images failed to identify landmarks with enough accuracy. Hence for our work, we have used the calibration approach from (Ferstl et al. 2015). We briefly explain the procedure used with this method in detail below.

The method detailed by Ferstl et al. (2015) as in Figure 5.8a is a user friendly and fully automatic calibration pipeline. Unlike a chessboard target, the procedure uses a circular target (Figure 5.8b) to capture a set of landmark images of both the image and depth sensor. This makes the procedure well suited for low resolution camera parameter estimation. As a first step, a set of images of the target are captured using the Hololens by walking around the printed target. During the calibration, the captured image and depth dataset of the target is then passed for marker identification. Following this, the circular markers in both the RGB and depth images are automatically identified and passed over for camera

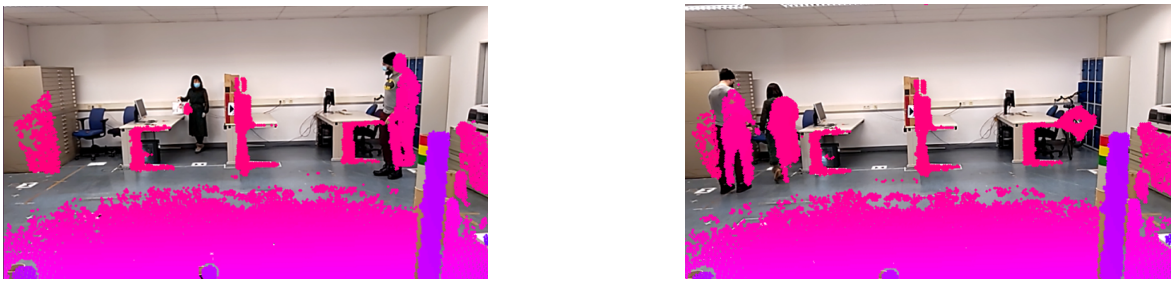
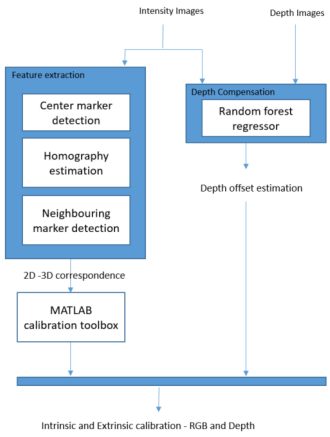
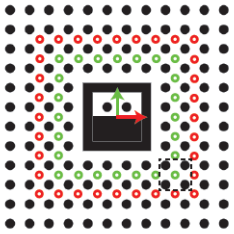


Figure 5.7: The Figure visually illustrates the errors in depth projection (highlighted with bright colours) using both the intrinsic and extrinsic parameters shared by Hololens research mode (Ungureanu et al. 2020).



(a) Calibration procedure detailed in Ferstl et al. 2015.



(b) Calibration Target in (Ferstl et al. 2015).

parameter estimation. The homography then computes the extrinsic registration between the two sensor. For this estimation, the Matlab calibration toolbox is used [5] as a part of this toolkit. Errors in depth sensing are then corrected in the depth compensation step. Consequently, the depth while capturing the circular target is compared to a learned depth model. The random forest regressor then estimates the offset in the captured and estimated depths which further improves the measurements of depth images. The calibration toolkit then outputs the intrinsic parameters of both image and depth sensor along with the extrinsic registration matrices.

3D Detection using MaskRCNN The MaskRCNN is a convolutional neural network that detects objects in images (as bounding boxes) and generates a high-quality segmentation mask for each detected instance. Hence to apply the network in localising pedestrians in 3D, the calibration parameters and the RGBD data from the Hololens are passed to the pipline with a MaskRCNN segmentation block as shown in Figure 5.9.

Given an RGB image with pedestrians in it, the image segmentation CNN returns the 2D bounding box, the segmentation mask and the probability score of each detected pedestrian in the scene. The segmentation mask then covers the complete contour of the detected person with all those pixel points that belong to the class *person*. Hence to identify the same person in 3D, all the depth points can be reverse projected using the obtained calibration

[5] https://www.mathworks.com/help/vision/camera-calibration.html

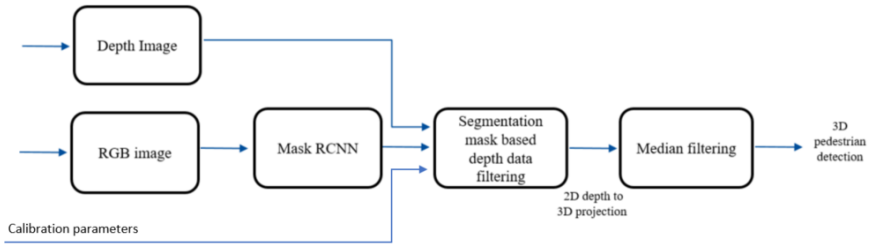


Figure 5.9: MaskRCNN based 3D detection using Hololens 2.

parameters. With this, all the 3D points that fall in the pedestrian mask are then identified as candidate points categorising them as pedestrians. A median filter is applied to the 3D candidates to remove artefact noises that would have been introduced due to dynamic projection errors (due to movement in the scene). As the possibility of such errors in influencing pedestrian localisation could be larger with increasing distance from the device, we only filter out noise points along the centroid of the detected mask and orthogonal to the camera. Hence our filtering only removes those points along the Z axis which might belong to the background of the scene. Once a filtered pedestrian pointcloud is obtained, each person can be approximated with a centroid and its enclosing box parameters *l, w and h.*

5.4 Perception- Particle Filter Tracking

Aimed at tracking multiple walking pedestrians in front of the Hololens, the particle filter implemented in this thesis takes the pedestrian detections as inputs from the detection module (either MaskRCNN or Frustum Pointnet) and tracks each of the localised persons as moving particles over the subsequent timestamps. Furthermore, to reduce the complexity of tracking in the 3D space, we strongly restrict the walking to the ground surface. For this, we projected each of the 3D detected persons onto the ground plane and represent persons with 2D positions (x_j, y_j). Each detected person is modelled as a particle filter that moves on this ground plane using a motion model. Then detections (or observations) from the subsequent timestamps are used to update the state of the filter based on the observation likelihood. The tracks for each detected person is managed by a track life cycle management that would update the filter based on how frequently the same persons are being detected in the 3D space. Once successfully tracked, each person is represented by a unique id and his/her track- *tracklets* that represents walking motion in front of the Hololens. The internals of our designed filter are further explained:

Particle Filter State: Each indoor moving person in front of the Hololens is represented using both the position and his/her moving velocity. Hence we represent pedestrian movement in our tracker using the state variable vectors k_j:

$$k_j : \{x_j, y_j, v_x, v_y\} \tag{5.2}$$

where v_x and v_y denote the velocity component in the x and y for a person walking with a motion velocity v.

Observation Likelihood Modelling: The observation likelihood would measure how well the target states $\{k_1, k_2, ..., k_m\}$ explain the observation data $\{d_1, d_2, ..., d_m\}$ from the detector. This function best describes the observational error in the input spaces over the tracking interval. At each time t, the likelihood that detection d_t to originate from tracklet k_i is :

$$p(d_t|f_k(k_i)) = exp(\sum_j -l_j(d_t|f_k(k_i))) \tag{5.3}$$

where function $l_j()$ compute the likelihood in euclidean space for the distance of our observations to the tracklet state.

Track Management: For each new observation, the state predictions for the filter are formed $p(k_t|d_{1:t-1})$ based on the previous tracked state $p(k_{t-1}|d_{1:t-1})$ and by moving the previous state using the motion model. As pedestrians move at relatively low speeds and along a uniform path, simple motion model like a constant velocity motion model could be used to represent the propagation of the particle filter state k_j.

Following the prediction step, associations for newly unassigned observations d_j and the Maximum a posteriori estimate of target track state k_t can be computed using the likelihood function. Based on the estimated likelihood, the particle filter states managed by the tracker undergoes a track lifecycle update *creation, update and deletion* based on the whether a new observation is associated to a tracked state or not.

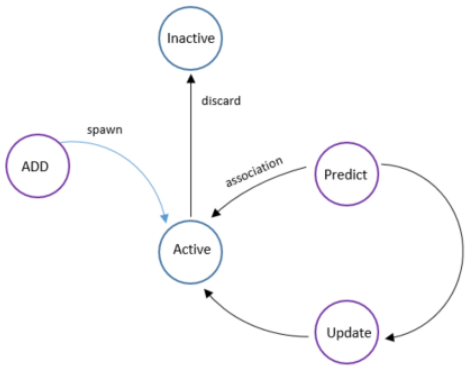


Figure 5.10: Different tracking life cycle states (in Blue) along with the state transitions of particle filter tracker (in Purple).

The Figure 5.10 depicts the different transitions of the particle filter tracker on the tracked states with every new observation at time t. The *Add* defines the creation of a new tracked object state k and the subsequent assignment of an Active track status to it. An active tracked state is propagated using a motion model in the *Predict* step. In the *Update* step, new observation data d_j is incorporated into each of the tracked target k_i following a successful association of tracked states with observation. With the *Update*, the filter state is updated based on the observation likelihood. All unassociated tracks move to the Inactive state and are discarded in a future point of time.

The tracker at each time stamp takes the 3D detection's as input and outputs the tracked position, heading angle and the track id of the person tracked over the time stamps.

6 Evaluating Pedestrian Scene Motion Perception

This section details with the evaluation procedure of how the scene motion pipeline discussed in the previous section has been evaluated. Section 6.1 details the procedure on how the F-Pointnet was trained and tested for detection with improvement on the Hololens data. The method discussed in the section is published as a research paper in (Kamalasanan et al. 2022a). Section 6.2 explains how MaskRCNN was implemented and its performance when detecting pedestrians in 3D in front of the AR device. Finally Section 6.3 discusses the implementation of the particle filter. Furthermore the tracking performance of using both F-Pointnet and MaskRCNN with the filter is elaborated.

6.1 Training and Testing F-Pointnet with 2D Human Pose

As the openly published F-Pointnet network was trained with the KITTI dataset (Geiger et al. 2012), we could not directly apply the pretrained model and its detection's for this thesis. Hence our method to evaluate the network for our pipeline involved retraining the network with RGBD data collected from the Hololens. For this we used the Simulated Shared Space Dataset (SSS Dataset) described in Section 4.1. Also rather than directly applying transfer learning, we first experimented with improving the feature space representation of F-Pointnet and evaluated it with respect to improved detection and technical feasibility for real-time detection. Finally we choose the trained model that best supported the constrain of real-time pedestrian detection for our AR motion influence.

2D Human Pose: A person standing in a scene with different orientations (e.g., back facing the camera, or titled walking posture) appears differently from a camera view point while captured. If the person is very close to the camera, his upper body and shoulder joints would be more visible compared to an image capturing him when further away from the RGB camera source. Also with the current improvements in the performance of image only human pose estimation algorithms (Toshev and Szegedy 2014, Wang et al. 2021) it might be possible to estimate 2D human poses with a high level of accuracy for a given image. Applying these poses estimated using images can be used to improve 3D detection algorithm like the F-Pointnet that also rely on RGB data in its pipeline. We hypothesise that the extra information from such image pose estimators could be used to improve the feature space representation for 3D pedestrian detection.

A state of the art 2D pose detection framework, OpenPose (Cao et al. 2017) for example can detect 25 landmark points of a human in the scene given the RGB image. These landmark points when interconnected for the shoulder, limbs and lower body parts (legs, feet and hip) would resemble the skeleton representation for the person. The extracted landmark points and the inter-joint pixel variations could represent significant 3D pose information for pedestrian perception.

While the baseline F-pointnet only extracted bounding boxes (with Yolo) of persons using images (as explained in Section 5.3.1), we attempted to further exploit the RGB data by

Figure 6.1: Openpose (Cao et al. 2017) based key-point detection with the dominant keypoints (S and H) that are considered for the Hand crafted features.

applying OpenPose joint estimates (as in Figure 6.1) and create hand crafted features from the poses. The hand crafted features are then used when training the F-Pointnet as as meta data to experiment improvements in 3D detection performance.

Hand Crafted Human Pose Features: Of the total 25 pose points estimated by the pose detector, we focused only on the dominant/ easily observable openpose key-points. This included the shoulder (keypoint 2,5), hip (keypoint 9,12), knee (keypoint 10,13), ankle (keypoint 11,14), elbow (keypoint 3,6), wrist (keypoint 4,7) and neck (keypoint 1). Also as the pixel representation for these points would not yield valuable information for a person at different distances from the camera, we craft features that were both scale invariant and 3D human pose representative. To make each of the features invariant to the effects of scale variations due to different distances from the camera, the features are normalised by a scale factor (SF). SF is the distance between the shoulder and hip joints, which is used for feature normalisation:

$$SF = |Joint_{hip} - Joint_{shoulder}| \tag{6.1}$$

We have developed the following feature representations of handcrafted features that would then be experimented by fusing with the F-Pointnet:

Distance Ratio (DR): The euclidean distance between the respective shoulder points and hip joints are represented in this feature.

$$S_N = \frac{|Keypoint_{(5)} - Keypoint_{(2)}|}{SF} \tag{6.2}$$

$$H_N = \frac{|Keypoint_{(12)} - Keypoint_{(9)}|}{SF} \tag{6.3}$$

As both the distance values are pixel representative distances of image coordinates, they are scale normalised using the factor obtained in the previous step. These features are then added as meta data to the network

$$F_{pose} = \{S_N, H_N\}$$

Optimised Distance Ratio (ODR): This feature representation optimises the distance ratio to better characterise feature points in representing human orientations. A smaller value of distance ratio correspond to people facing the camera from the side; while a larger value might meant that they were captured with their front or backs oriented towards the camera. Hence to amplify the orientation data that were represented as joint distances, log transformations were applied. This operation boosted the smaller shoulder and hip lengths while also suppressing excessive dominance of front-facing or back posed people in the scene.

$$F_{pose} = \{-log(S_N), -log(H_N)\}$$

Optimised Distance with Keypoint Position and Distances (ODPD): This feature encoded a more detailed summarisation of the 2D tilt and turns of persons in the scene. For this along with ODR features, all the normalised joint positions and distances from the pose estimator were included into the feature space.

Normalised position (N_p) as represented in (Li et al. 2020) was obtained by first transforming the image pixels to an interim coordinates where the neck was the origin. The arm (2-7) and legs (9-14) points were then normalised by the scale factor SF. For the legs, four distance features would include the distance between left hip and left knee, left knee and left ankle, and the same for the right legs. While for the arms, the four features include distances between the left elbow and left shoulder, left elbow and left wrist, and corresponding features from the right arm.

In total the eight position and distance values in scale agnostic pixels were then concatenated along with OPR to represent ODPD. The choice of position and length of joints was inspired from (Li et al. 2020) where motion estimations of walking person were approximated from 2D images poses of human joints viewed from on-board camera of vehicles. We hypothesised that including all joints would be beneficial in situations when a person is only partially visible to the camera. In such cases the computation of ODR would fail.

$$F_{pose} = \{-log(S_N), -log(H_N), N_p, N_d\}$$

In the training phase, the *Baseline F-Pointnet* and its variant with pose features - *F-Pointnet with 2D Pose* were investigated. The Simulated Shared Space dataset (SSS Dataset) was used as the training data with 3D pedestrian groundtruth created using the semi-automated labelling method (see Section 4.1).

Baseline F-Pointnet: For the Baseline network, the implementation by (Qi et al. 2018) for LIDAR sensor was adapted to the Hololens and its 3D coordinate system. Also the predefined 3D anchor boxes for class *pedestrians* were tuned to match the average pedestrian dimensions (l,w,h) of the SSS Dataset.

F-Pointnet with 2D Pose: The Figure 6.2 depicts the two modifications that were made to fuse the hand crafted pose features with the baseline F-Pointnet. Firstly as 2D human pose features had to also be extracted from the RGB images, along with detecting pedestrians in the first stage of the F-pointnet pipeline, we also passed the images to a pose detector. The three variants of high level pose features F_{pose} were then computed for a features fusion step. Secondly as the extracted features (F_{pose}) had to be combined with pointnet based 3D features, we choose the final stage of the F-Pointnet for feature fusion. The choose of this fusion point was to maximise the contribution of the 2D pose data to the feature space and to ensure its significant impact in the pose regression stage that followed the fusion point.

With the discussed enhancements to the baseline variant, we experimented on how the three variants of high level features impacted a baseline F-Pointnet detection accuracy. For the experimental evaluation, we used the state-of-the-art pre-trained models. We used YOLOv3 pretrained on COCO for image detections and OpenPose for the pose extraction. With the detected 2D poses from OpenPose, handcrafted features *Distance Ratio* , *Optimised Distance Ratio* and *Optimised Distance with Keypoint Position and Distances* were computed. Then evaluations were made to study improvements to pedestrian detection performance when using these extra features.

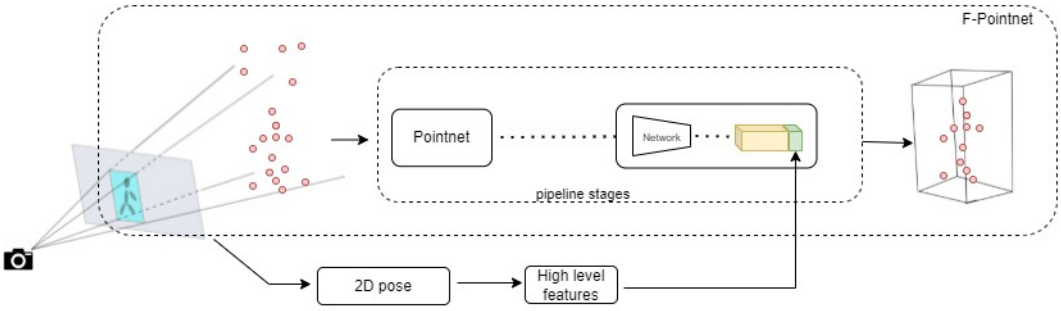


Figure 6.2: High level representation of hand crafted pose features fused with the F-Pointnet.

Results and Discussion We trained both *Baseline* and *F-Pointnet with 2D Pose* for 150 epochs with a batch size of 32. We completed the training on Nvidia 1080Ti GPU machine with the dataset randomly split into training (80 %) and test sets (20 %). We measured the performance of the network with the AP and AOS indicators. While the former was used to benchmark improvements in 3D object detection, the later to evaluate whether orientation estimations of detected people were better when 2D pose was added.

Methods	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$	AOS
Baseline F-PointNet	**0.8910**	0.4957	0.0108	0.6596
F-Pointnet with 2D Pose [Distance Ratio (DR)]	0.8770	0.5004	0.0303	0.5878
F-Pointnet with 2D Pose [Optimized DR]	0.8688	**0.6470**	**0.0660**	**0.7477**
F-Pointnet with 2D Pose [Optimised DR w/ P&D]	0.8093	0.6358	0.0587	0.6599

Table 6.1: *F-Pointnet with 2D Pose* with alternative feature selection using high level pose information. The scores obtained for the baseline are highlighted for comparision with the others.

The Table 6.1 shows the AP performance for different IoU thresholds (0.3, 0.5 and 0.7) along with AOS scores that were estimated for both the baseline and F-Pointnet improved with hand crafted features.

It can be noted that the baseline network performed relatively well for low IoU threshold (IoU = 0.3). However when higher IoU thresholds (0.5 and 0.7) were applied to understand performance with larger overlaps between the estimated and the real 3D shape and position of persons, the F-Pointnet with improved feature representation performed better. At both IoU 0.5 and 0.7 the network that was concatenated with Optimised DR features performed higher. As for the AOS score measured to see orientation improvements, even when adding pose features indicated towards advancing correctly predicting how persons rotated their body; these improvements when compared to the baseline were not significantly high. The baseline improved by only 13% when ODR pose features were included with input RGBD data.

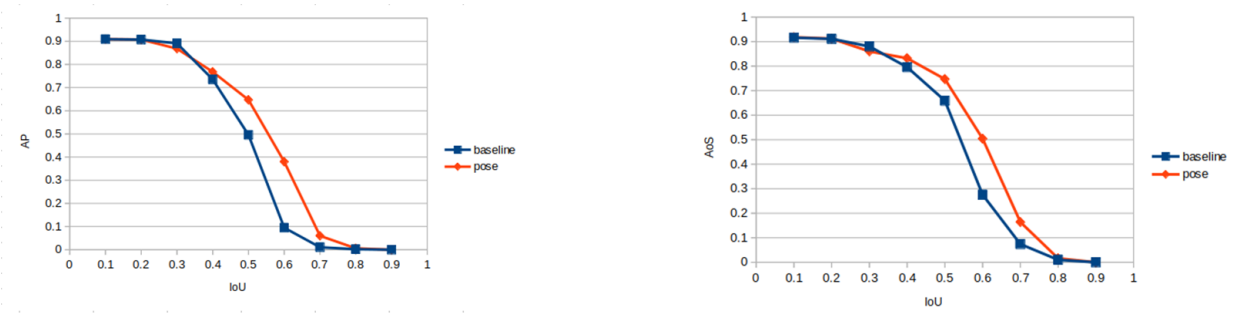


Figure 6.3: The AP (left) and AOS (right) for different values of IoU threshold for ODR compared against Baseline F-PointNet.

The Figure 6.3 shows the baseline compared against the best performing hand crafted feature (OPD) for both the 3D detection performance and the corresponding orientations estimations. It can be seen that the AP and AOS both seemed to perform relatively the same as the baseline for low IoU thresholds (< 0.4). However as the IoU threshold increased, even when pose features contributed towards improving both the detection and orientation estimation (indicated by higher AP and AoS values), the scores did not show considerable improvement for all values of IoU.

Feasibility of Applying 3D Detection's with 2D Pose: While the above experiments have focused on training a baseline F-Pointnet network with Hololens data and testing improvement using pose features, the results indicated higher performance in both 3D detection and orientation estimation for only certain IoU thresholds. Better performances were noted for only higher IoU (>0.4) overlaps between detection's and ground-truth. As our work is focused on using a detection by tracking based motion pipeline, we hypothesis that lower positioning and orientation errors can be improved in the tracking stage of our pipeline.

Hence using the lower IoU threshold (IoU = 0.3) as a reference, we implemented the *baseline F-Pointnet* and its modified *F-Pointnet with 2D Pose* using the code released with (Martin-Martin et al. 2021). While the chosen Yolo v3 was supported by ROS [1], a similar compatible

[1] *https : //github.com/leggedrobotics/darknet*$_r os$

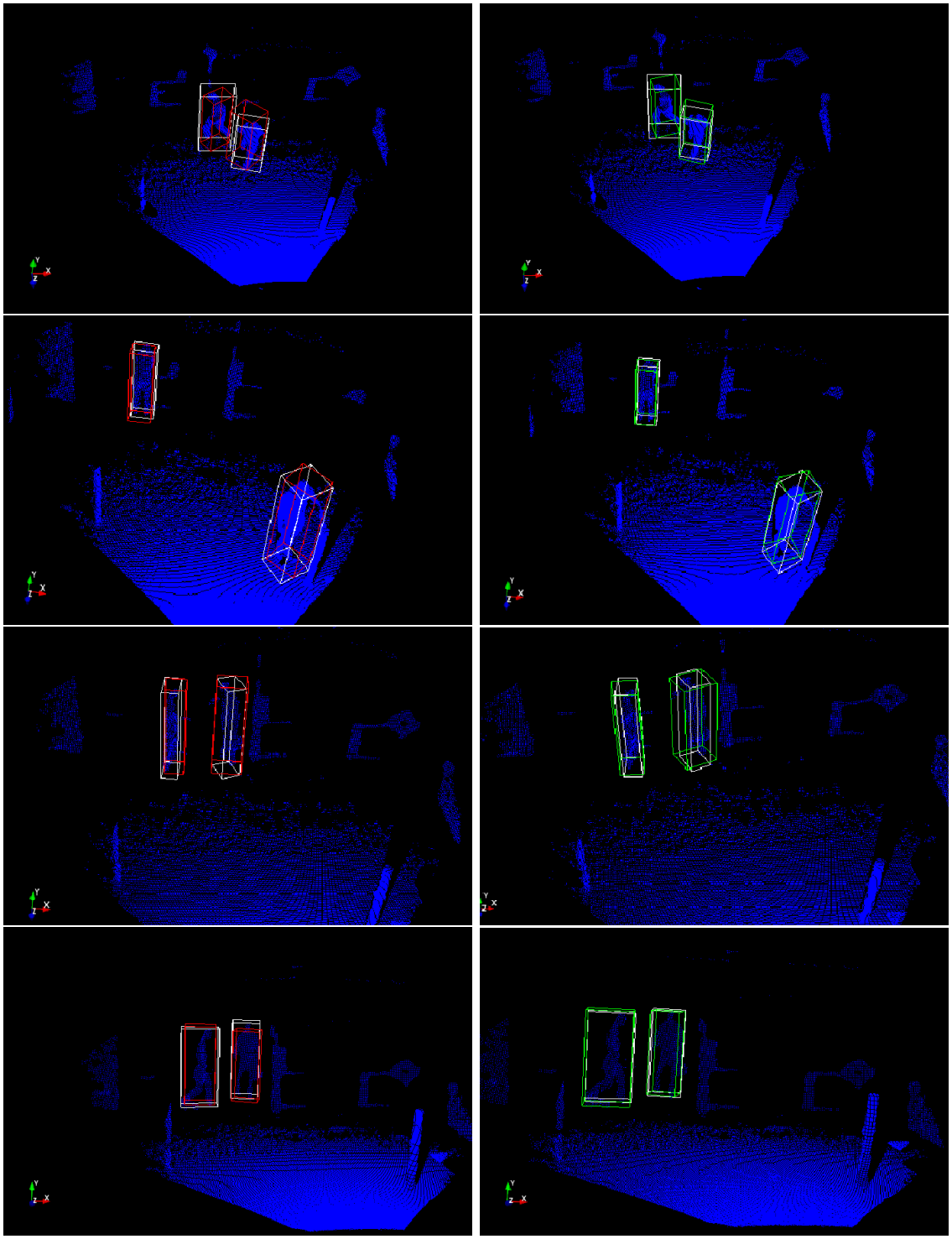


Figure 6.4: Qualitative comparison of pedestrian 3D detection results using baseline (red bounding boxes on the left) and our proposed approach using ODR features (green bounding boxes on the right). The white bounding boxes are the manually annotated ground truth.

OpenPose implementation [2] was used along with it while testing. When the pose modified version of F-Pointnet was tested online on Nvidia Tesla X GPU, larger processing time (>10ms per RGB frame) for OpenPose was noted. This severely limited the real time execution capability of F-Pointnet. Hence the baseline F-Pointnet with higher frame rates ($> 10fps$) is choosen for the rest this thesis to influence walking with AR perception.

6.2 Testing MaskRCNN for 3D Detection

To apply MaskRCNN 3D detection to the Hololens 2, we first calibrate the RGBD sensors of the device using the procedure by Ferstl et al. (2015) as mentioned in section 5.3.2. For this, we print the calibration target (80cm x 80 cm) and create a dataset of RGB and depth intensity images with the device positioned differently capturing the calibration target. We then approximate the RGB and depth camera intrinsic along with the registration between the sensors using the Matlab code implementation [3] released with the algorithm.

The Figure 6.5 shows the visual results of the calibration experiment with the calibrated depth projected (in blue) onto the RGB images. As noted in the images, it can be observed that the depth points for the objects in the scene correctly fall onto its corresponding RGB pixels. The incomplete depth projections in the upper right and left corners are mainly due to the limited field of view of the Hololens depth sensor. A quantitative evaluation of the calibration accuracy is currently not covered in the scope of this work.

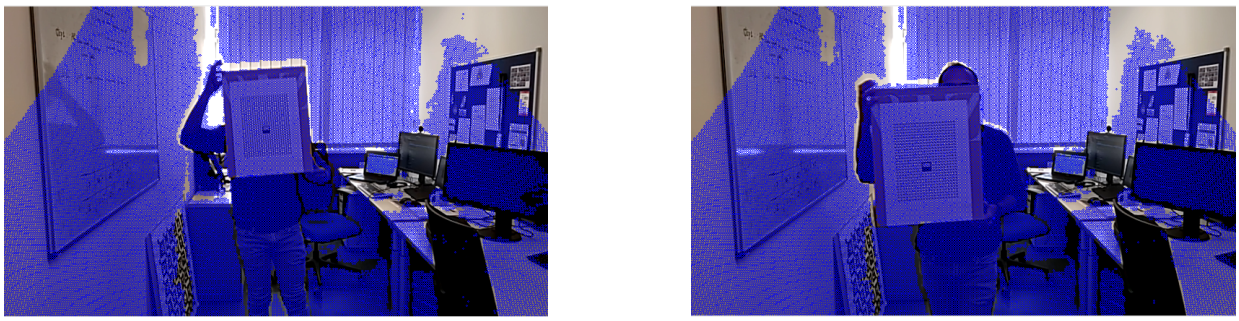


Figure 6.5: Figure on the left and right show the results of applying RGBD Calibration (Ferstl et al. 2015) to the Hololens. The discontinuity of point-cloud projection to the upper corners is attributed to the limited field of the depth sensor.

MaskRCNN Implementation To achieve pedestrian detection in 3D, we have used the OpenCV MaskRCNN implementation [4] that was used to detect objects using a pretrained COCO model. We refrain from retraining the model using the Hololens data to establish performance baselines using the default settings. Of the objects detected by the model, we selectively filtered out all objects belonging to the COCO class *person* while applying a object probability threshold (p =0.3) as the minimum confidence interval.

For each detected person and the maskRCNN segmentation mask, we used the 3D pointcloud from the Hololens to localise the detected persons in 3D. For this all those depths points that

[2] $https://github.com/ravijo/ros_openpose$
[3] $https://github.com/RobVisLab/camera_calibration$
[4] https://learnopencv.com/deep-learning-based-object-detection-and-instance-segmentation-using-mask-rcnn-in-opencv-python-c/

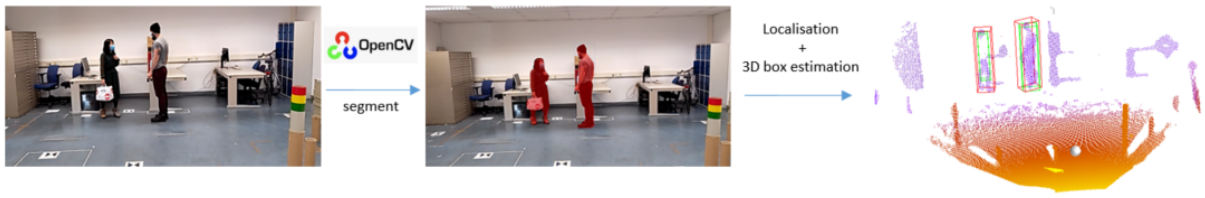


Figure 6.6: The pipeline implementation for MaskRCNN where raw RGB image, depth and the masks(in red) from the opencv segmentation are used to estimate the 3D bounding box for pedestrians in the scene.

belong to the segmentation mask *person* were filtered to obtain the pedestrian pointcloud points. Following the filtering step, we estimate the position and the dimensions of the enclosing pedestrian box in 3D from the person cloud. For this, we compute the centroid of each filtered cloud instance and the estimate the enclosing 3D box (*length*, *width* and *height*) based on the min and max of the pointcloud along the x, y and z axis respectively. The Figure 6.6 illustrates the complete 3D pedestrian detection when applied to the hololens data.

Accuracy Estimation and Results To estimate the accuracy in correctly identifying and localising pedestrians in 3D using the MaskRCNN segmentation, we apply both quantitative and qualitative evaluation in this thesis. For the quantitative assessment, we estimate the Average precision (AP) for the performance of this pedestrian detection on the Simulated Shared Space dataset (Section 4.1). Since the 3D detection output returned from our current algorithm only predicted 3D boxes (c_x, c_y, c_z, l, w, h) without orientation, the SSS dataset ground-truth created using semi-automated labelling could not be used as in Section 6.1. Hence to create reference labels with orientation in 3D, a small part of the SSS dataset (n= 125 sequences) was manually annotated as ground truth in our AP estimation. Also for the manual annotation, the Labelcloud (Sager et al. 2021) was used.

For the estimation of AP, we only focused to compute the precision metric for lower pedestrian bounding box overlaps with its ground truth (IoU = 0.3). Applying this IoU threshold, the pedestrian detection's from the 3D MaskRCNN pipeline resulted in an AP of 0.67. Furthermore to make qualitative inferences, we visually plot both the results from the detector (in blue) and its groundtruth (in white) as shown in Figure 6.7. As noted in the figures, it can be observed that the detector successfully identified pedestrians even when their point representation in 3D were sparse (as in the top row). Furthermore as the 3D detection was mainly based on the segmentation RGB image mask, the field of view of the RGB camera and the image pixels of the person obtained for each image influenced the dimensions of the estimated 3D bounding box. This meant that if the RGB camera captured a person partially, then all the depth points of the person missed by the image sensor were not considered resulting in partially correct bounding boxes. This effect was more pronounced when the person stood closer to the RGB camera of the Hololens. Lastly it can be noted that not all 3D positions correctly overlapped with the ground-truth, this we hypothesise could be largely due to the time synchronisation error between the RGB and its corresponding depth image and also due to inherent errors in the detection algorithm which needs to be investigated in the future.

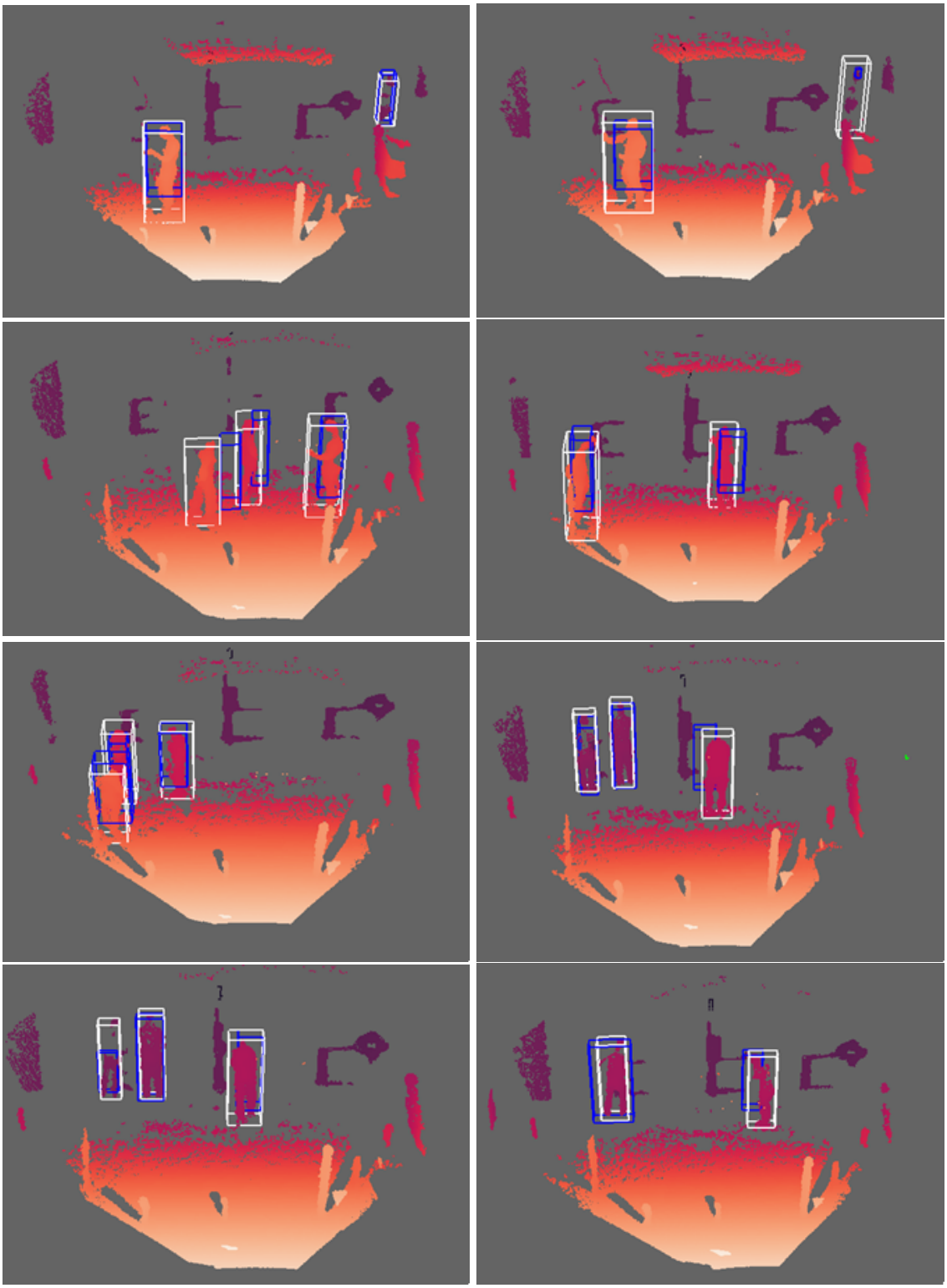


Figure 6.7: The Qualitative comparison of the pedestrian 3D detection when comparing the groundtruth (highlighted in white) to the detection from the Maskrcnn detection (indicated in blue).

Real Time 3D Pedestrian Detection with MaskRCNN As the primary objective of this thesis has been to prove the motion perception capability of a pedestrian pipeline for AR, the pedestrian detection approaches with MaskRCNN was implemented with OpenCV for online 3D pedestrain detection. The pipeline for MaskRCNN was implemented at run-time using C++ with GPU support for both quick scene segmentation and inference. The method achieved a frame rate of 10 fps for GPU inference on Nvidia Tesla X.

6.3 Pedestrian Tracking and Performance

In this section, we explain the particle filter implementation and how pedestrians were tracked in 2D using the results from both detection methods- the **Frustum Pointnet** and **MaskRCNN** detection. We choose the Bayesian particle filter due to its ability to represent nonlinear and non gaussian systems using multimodal representation. This then helps to approximate position uncertainty of pedestrians when tracked using either of the detection methods. Lastly we explain how we computed the tracking performance for both detection approaches using the MOT metric.

Our tracking algorithm is based on estimating the position and orientation of each target pedestrian by a particle filter. For this we use the output of the 3D detector as the observations; and estimate the time-evolving posterior distribution of persons location using a weighted set of particles. The state x ={x, y, z, ϕ , v } of our tracker consists of 3D position (x, y, z), orientation (ϕ) and velocity v. However in the scope of our current work, we assume the person to be walking on the ground plane and hence take z to remain a constant.

Particle Initialisation :For every pedestrian detected by the 3D detector, the algorithm initialises a tracker individually. The particle positions are sampled from a normal distribution centered at the positions estimated by the 3D detector. To initialise the velocity of the particles, we followed a behavioural motion model based initialisation as in (Dimitrievski et al. 2019). For this we use the IKG Pedestrian Tracking dataset (section 4.2) and experimentally fit a distribution for the velocity profile for the moving pedestrians.

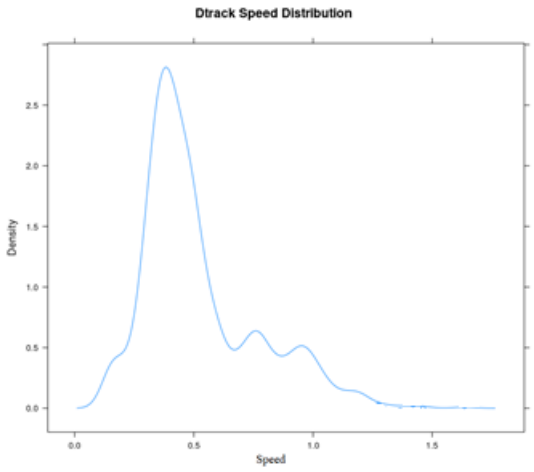


Figure 6.8: Speed density plot for persons moving in the IKG Pedestrian Tracking Dataset.

We observed that the distribution for walking speeds for persons follow a truncated multi-modal gaussian distribution as shown in Figure 6.8. The distribution could be summarised as:

$$p(|v|) = \sum_{i}^{n} a_i N(||v||; \mu_i, \sigma_i) \qquad \text{if } v \in [0, 1.2] \tag{6.4}$$

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(x-\mu)^2}{2\sigma^2}) \tag{6.5}$$

The mean values from the distribution indicate that the people in the scene for the IKG Tracking dataset were mostly walking at slow walking speed. The gaussian components for the multi-modal distribution being $a_1 = 0.6$, $\mu_1 = 0.36$, $\sigma_1 = 0.036$ for the first , $a_2 = 0.6$, $\mu_2 = 0.40$, $\sigma_2 = 0.111$ for the second and $a_3 = 0.6$, $\mu_3 = 0.82$, $\sigma_3 = 2.182$ for the third components respectively. The components were approximated using R programming.

Motion Model and Tracker Position: In the filter, the particles were propagated using a constant velocity motion model :

$$\begin{aligned} x_k &= x_{k-1} + x_{dot} * \Delta t \\ y_k &= y_{k-1} + y_{dot} * \Delta t \end{aligned} \tag{6.6}$$

$$\begin{aligned} x_{dot} &= v * cos(\phi) \\ y_{dot} &= v * sin(\phi) \end{aligned} \tag{6.7}$$

For every time t, we compute the longitudinal and lateral velocity (Equation 6.7) and move the particles on the ground plane based one the previous location of the particles and the orientation estimated by the tracker.

Although represented by a probabilistic estimate (using a set of particles), the position of the tracker in each time stamp can be estimated using the strongest mode of the kernel density estimate of the particle spread. To compute the mode, we project the particles of the filter onto equally spaced (N=100) 2D grids. The grid size would hence depend on the spread on the particles propagated during prediction at time *t*. For each of the grids, we fit a 2D Gaussian and identify the local maxima for the 2D grid (Figure 6.9). The maxima approximates the position which then updates the tracker state to represent the position of the person in the scene.

Data Association: Following the particle prediction steps, to decide on which incoming new detection at time *t+1* should guide the tracker estimate for *t*, we solve a data association problem assigning atmost one detection to each of the tracked objects. As there might be cases of one-to-many or many-to-one associations between the two sets, we use the Hungarian algorithm to optimise the problem of assigning N tracker targets to M detection's.

The matching using the Hungarian works as follows: A cost matrix **K** is first computed for each pair (n,m) modeling the affinity between the detection's and tracked positions. The

Figure 6.9: The figure on the left shows the particle spread of the persons propagated during prediction while the picture on the right shows the position estimated using Gaussian mixture modelling.

affinity cost matrix is computed using the euclidean distance between the detection and tracker positions in the scope of the current work.

The 2D matrix would then be characterised with each row represented a distance cost computed for each of the tracked states. In the hungarian optimisation step, each of the columns is reduced by the lowest cost along the vertical axis. With this step, if the number lines drawn to cover all of the zeros is the matrix match the order of the matrix, then an optimal assignment is expected to be reached otherwise the procedure is repeated. Once the affinity matrix is optimised, each of the zero entries in the cost would represent an optimal assignment.

Observation Model : With every positive association, we compute and update the weights of the particle filter estimating and updating the conditional likelihood of the new observation given the propagated particle.

$$w_{m,p} = p(y_t|x_t) = I(n).p_n(\frac{p-d}{\alpha}) \tag{6.8}$$

For the computation of likelihood we use the distance between the particle p and the positively associated detection d evaluated over a normal distribution (p_n) as in Equation 6.8. The term I(n) represented a association function that returns 1 when the detection n was associated with the tracker particle m and zero otherwise. The resulting set of weighted particles would then approximate the posterior distribution of the latent state space. During the resampling step, high weights are duplicated while weights with low values are discarded resulting in a set of uniformly weighted particles. We follow importance sampling as detailed in (Djuric et al. 2003) as the sampling procedure for our filter.

Track Management : As tracked pedestrian leaves the scene or when new persons enter the scene a lifecycle module to manage the tracker and its tracked id was implemented for the particle filter. For this, we consider all unmatched detections $D_{unmatch}$ from the data association step as potentially new persons in the scene. Also to avoid false positive trajectories, a new id was not assigned until $D^p_{unmatch}$ the person has been continuously matched for p frames. Also to prevent inactive states from remaining in the scene, we kept track of each unmatched tracked person in the scene for the different time stamp and

discard them for tracked duration > Age_{max} . In our work both p and Age_{max} are set to three. Finally , to evaluate the performance of the particle filter tracking on the two detection based approach, we implement the bayesian filter in C++ and quantitatively & qualitatively evaluate it using the IKG Pedestrian tracking dataset.

Experiments with IKG Tracking Dataset: We perform a comparative study where the results from both the detection approaches are measured visually and using MOT tracking metrics. For this, we evaluate the tracking algorithm on IKG Tracking Dataset for sequences that contains two pedestrians walking in the scene in front of the Hololens device. For the sequences that matched this requirement, we use the RGB and 3D information from the Hololens and created detection data for comparisons. For the F-Pointnet detection, we run the detection with GPU inference and record the detection tracks at 12 fps. The semantic segmentation based MaskRCNN 3D detection is run in real-time recording the detected person in the scene at 10 fps.

To make visual observations of the tracking performance, we created birds eye view(BeV) images of the captured point cloud scene of tracking dataset. We then draw the detected tracks and the corresponding track id assigned to each of the detections for different time stamp as shown in Figure 6.10 and Figure 6.11. For illustration, each tracked person in the image is indicated using a short line indicator. The location of the indicator then shows his/her tracked position and the tracked heading is indicated by its angle. The heading orientation shown in each image is the angle the line makes with the vertical axis 3D axis projected to 2D. To further describe the walking sequence shown in figure 6.10, a single pedestrian first enters the scene (row 1) and continues to walk in the field of view of the Hololens while a second person joins him in the scene (row two). Both persons walk along independent motion path for the rest of the sequence. There are also points when they cross each other or walk close to the other person as detailed in section 4.2. Empty point cloud patches (frustums) can be observed in the BeV images as pedestrians walk in front of the 3D sensor. This is due to pedestrians walking close to the device and occluding all objects and points in the behind.

As it can be noted in both the figures for tracking, the tracks from the F-Pointnet were more prone to tracking id switches and discontinuous tracks when persons crossed each other or when persons were walking away from the Hololens. The tracks from the MaskRCNN performed comparatively better for all the cases that were studied. Also looking at the different ID that were assigned throughout the sequence, the number of ID switches were very high for the F-Pointnet based detection pipeline. The MaskRCNN pipeline performed well even when persons were further away from the Hololens or when moving very close to the device. An offline analysis of the F-pointnet track noted multiple missed and false 3D detection to be the primary reason for the poor performance of the Frustum based pipeline.

Tracker Evaluation To quantitatively evaluate the tracking performance from the two approaches F-Pointent and MaskRCNN. We used the tracking results along with the groundtruth from the IKG Pedestrian Tracking dataset and calculated the MOT metric. Furthermore as the ground truth data was captured using an external tracking system (section 4.2) a post processing groundtruth alignment step was added in our evaluation to obtain accurate ground truth for the tracking sequences. In this step, the groundtruth data from the external optical tracking was matched to the Hololens coordinate system. Also the data

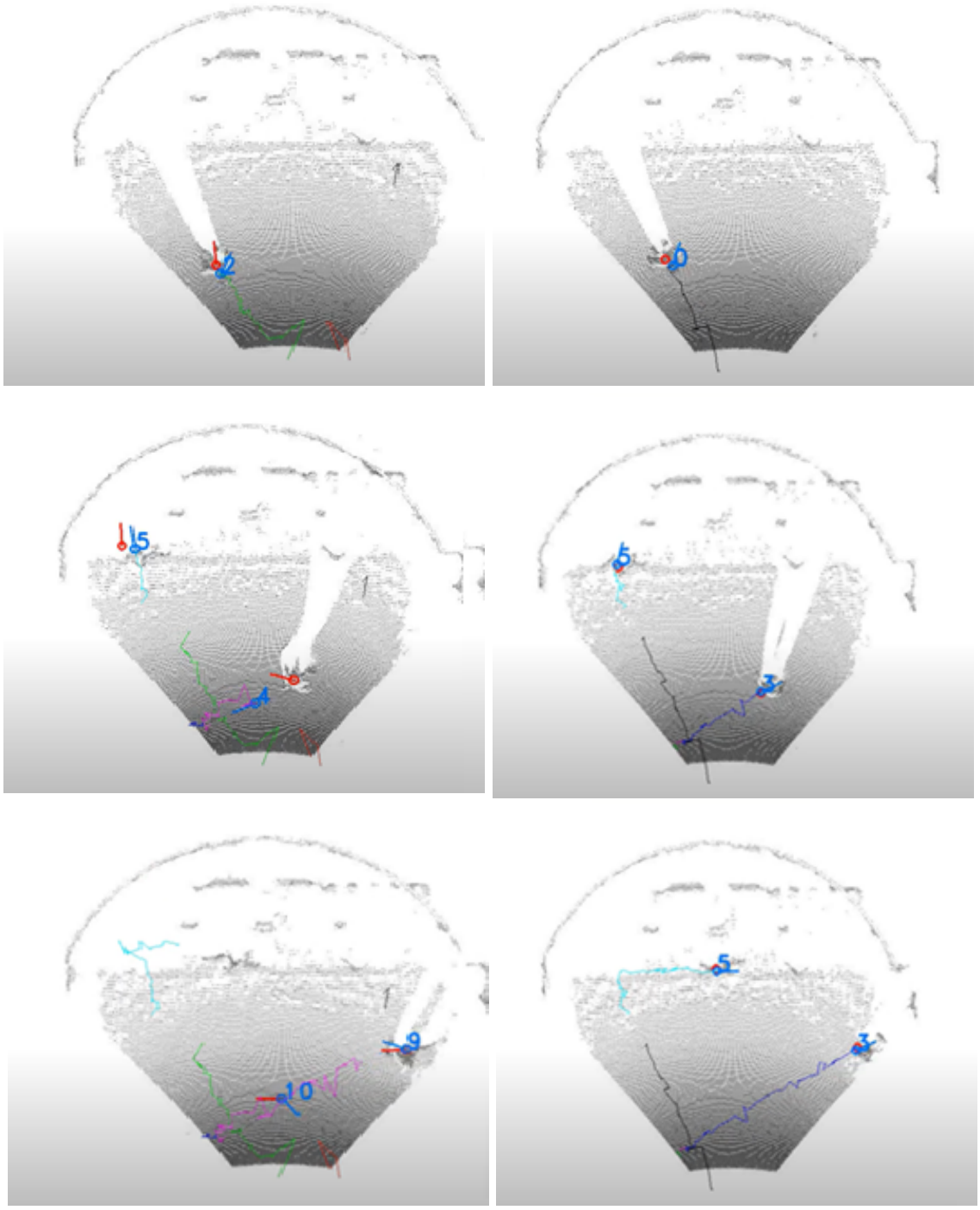


Figure 6.10: The figures shows a visual comparison of the tracking results from a point cloud birds-eye-view. The scenes represent a single person walking into the field of view of the device (row 1) and a second person crossing his walked path from behind (row 2 and row 3). The left column show the pedestrian tracks and ids (number in blue) using the F-Pointnet while, the right column show tracks and ids from MaskRCNN detect and track pipeline respectively.

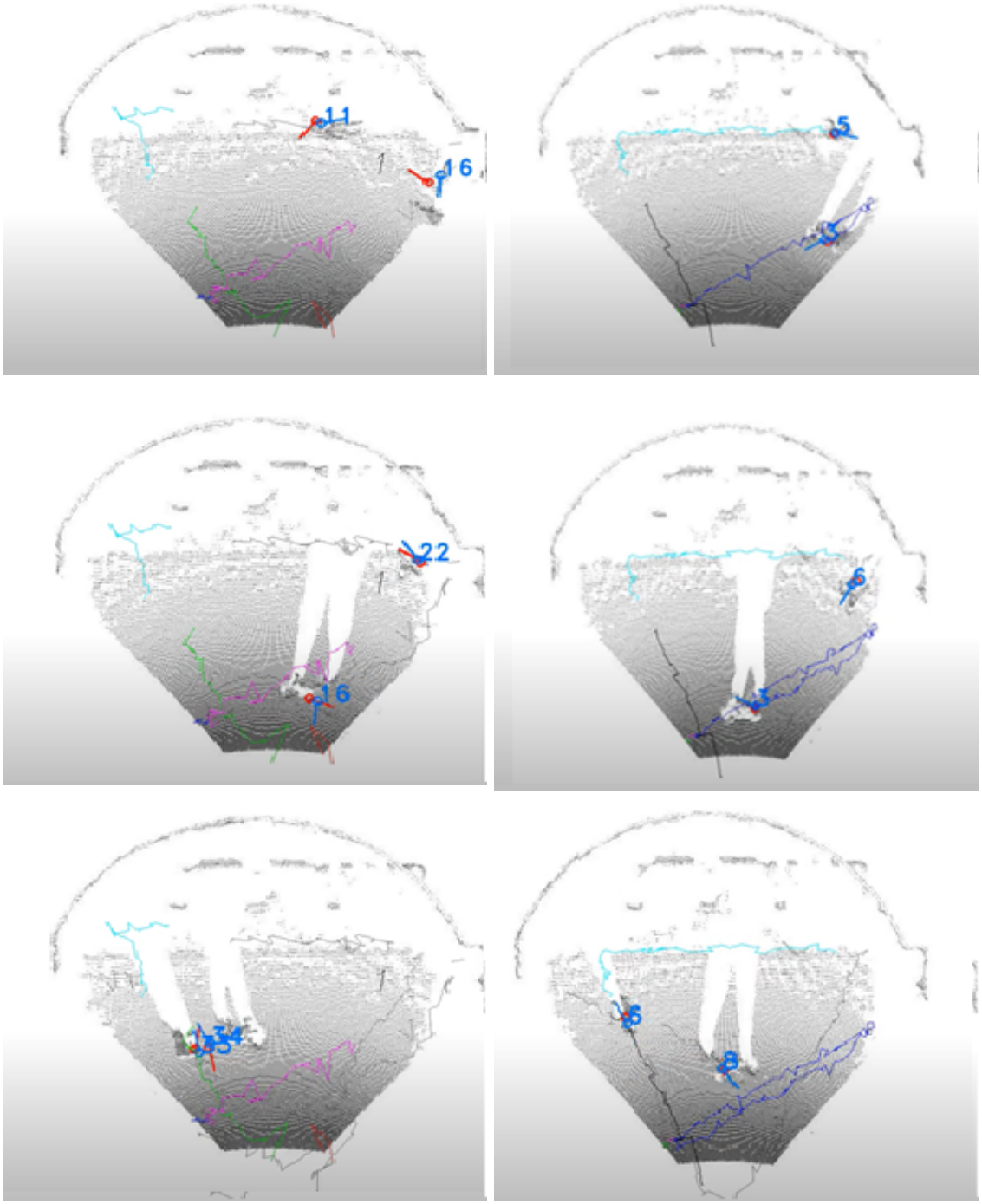


Figure 6.11: The Figure shows the visuals of subsequent motion for the two persons as the move in triangular walking pattern within the Hololens 3D field of view. The different coloured paths indicate id switches as a tracker fails to detect the same persons correctly. The tracked ids number for detection's from F-pointnet (right column) are considerebly higher than the MaskRCNN (left column) tracked persons.

was temporally aligned to match the detections from the device for both the approaches. To spatially align the coordinate system of the ground truth data with the data from the AR device, we apply an ICP alignement using Open3D [5]. Temporal alignment is achieved by first downsampling the data to match the frame rates from the Hololens detections, followed by a manual time offset correction.

Following the alignment, we computed the multi-object tracking (MOT) metric to evaluate pedestrian tracking performance. The Table 6.2 illustrates the performance of the tracker for both detection approaches.

Table 6.2: CLEAR MOT tracking accuracy.

Approach	MOTA	MOTP
F-Pointnet	71.29%	0.684
MaskRCNN	81.39%	0.878

Results The MaskRCNN pipeline clearly performed above the F-Pointnet in both precisely localising pedestrians (based on the MOTP score) and in following each tracked person (ID switches) as indicated by the higher value of MOTA. The error in tracking pedestrians using the MaskRCNN Pipeline was as low as 20 centimeter as evaluated with this metric. Hence pedestrians amongst other objects, could be identified and tracked in ego-view with such low errors when walking in front of the Hololens AR headset using the developed pipeline.

[5] `http://www.open3d.org/docs/release/tutorial/pipelines/icp_registration.html`

7 Influencing Behaviour by Visualising Future Motion

This section details on how visualisations (applied with an AR interface) of the surrounding pedestrian motion influences walking path choices of an ego-user. For this, we visualise the future path of nearby persons with AR and then study how path choices would differ when seeing future paths with different types of 3D representations. While designing an algorithm to predict pedestrian future motion (Zhang et al. 2023) is currently not covered in this thesis, the motion influence that result when visualising predicted future paths is studied. We propose 3D designs to represent predicted future trajectories and also investigate whether visualising motion prediction accuracy or uncertainty plays a role in path choices. In the following sections, along with our AR future trajectory study, we explain the terms *scene* and *prediction complexity* that are used in the work. Also, the different visualisations and what they represent are explained. Furthermore, the different hypotheses that were formulated to design the study are explained before the user study section.

Scene Complexity Every indoor or outdoor scene with moving objects/persons tends to exhibit an inherent level of complexity for a viewer when estimating future events. The level of complexity would go up for a task of prediction walking steps if the space is large and the movement of other persons are completely random. Hence the availability of free space, the existence of objects on the walking path or movement nearby could influence the path a person plans to take. The complexity of navigating a scene due to the availability of walking space and external factors(other persons) is addressed by this term in our work.

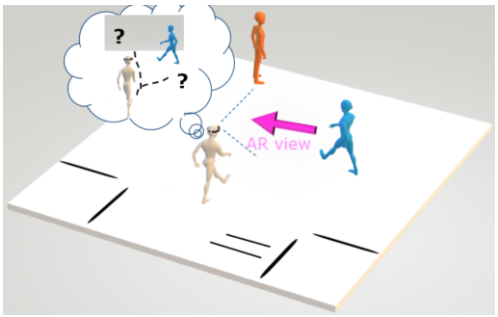


Figure 7.1: Future path visualisation which could trigger different path choice behaviours.

Prediction Complexity in the the context of this thesis study refers to the level of difficulty to guess future paths for a given scene. For this, we categorise complexity based on how visual clues available in a scene are used to guess the future steps as:

- Low Complexity: Observing another person's motion would suffice to estimate their future path.
- Medium Complexity: Visual observation of motion, combined with inferences from their body language (e.g., hand movements, foot positioning), might be needed to predict future actions.

– High Complexity: It might be hard to make an estimate of the future path without the aid of the future visualisation for the person in the scene (e.g., when a person would turn at a corner after stopping).

Visualisation of Future Path Having observed a few walked steps, the future path a person might take is an approximation of what could be his/her next possible steps in the upcoming point of time. As an estimate is always a nondeterministic quantity, the predicted path is approximated with statistical probability that could either have a large uncertainty or high confidence. Within the scope of this work, we categorise the AR future visualisation as:

– Simple AR predictions: An AR view that would indicate the future position of others based on where they would walk next using 3D symbols (like a 3D arrow).

– Informative AR predictions: The visualisation in AR that would not just show the future positions, but also encode some qualitative information (e.g., confidence/ uncertainty) of the estimate. This is visually represented by varying the appearance of the above mentioned simple AR predictions.

For the study to evaluate the effects of seeing future path from safety perspective, while designing the future visualisation itself is important, another significant aspect is the study design. In this thesis, we first formulated a set of hypothesis on how people might react to future visualisation and then designed our study to test the hypothesis. The results from the study were used to evaluate if the hypothese created were true or false. The following were the hypothese:

H1 - Safety: *When the walking path of a Hololens user would conflict that of an another person, the user decides to walk closer to the conflict point when seeing future paths explicitly using an AR medium.* This hypothesis is formulated based on observations made in pedestrian collision avoidance. People tend to choose a larger safety margin to avoid a collision in natural walking condition to avoid bumping. However when AR explicitly communicates motion of others, people might optimise walking detours as they feel more safe with extra augmented information. This can then result in shorter and closer paths to the conflict point.The Figure 7.2 illustrates how the Hololens user sees the virtual path in AR view of the crossing pedestrian. Then deciding to walk a shorter and closer path (in black) against his natural avoidance path (in blue).

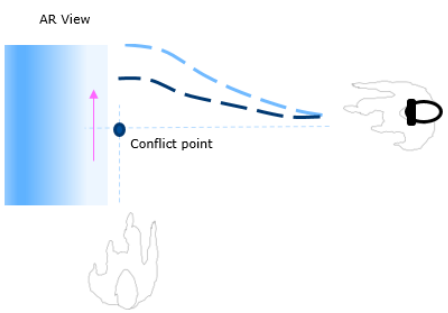
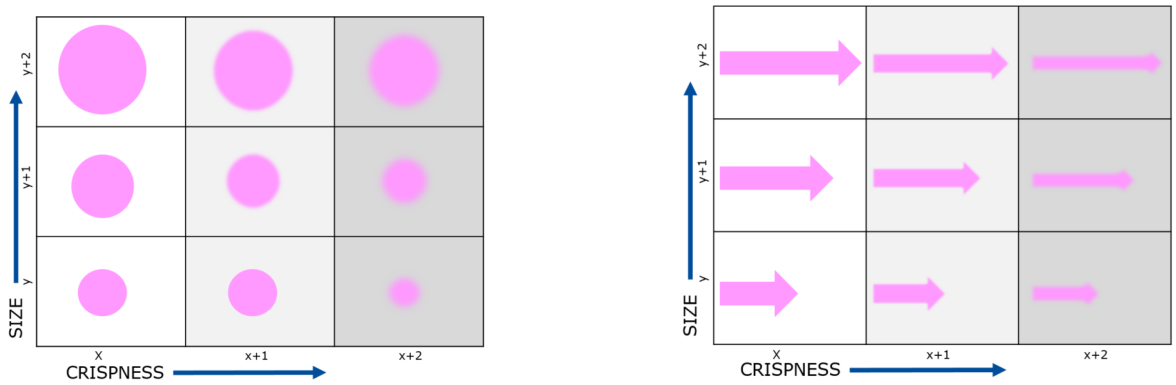


Figure 7.2: Path choice (as per the hypothesis) for a hololens person (right) who sees the future path (indicated in an arrow) of others in AR and decides to walk closer to the collision spot.

H2 - Appearance: *Visual representations of future position with informative AR could prompt different walking behaviours based on what information is conveyed.* This hypothesis is formulated based on a recent work (Fuest et al. 2023) that tested the effects of visually varying map symbols for navigation. In the work, based on the pollution level each symbol was encoded with a different variant (Roth 2017) like size, color e.t.c.,. This was then observed to prompted drivers opting for cleaner routes. Using visual variables size and crispness from (Roth 2017), we represent confident and uncertain future path respectively.



(a) Conjugation of variables for circles

(b) Conjugation of variables for arrow

Figure 7.3: Applying a conjugation of length and crispness variables to symbols (circle and arrow) as in Roth 2017.

The concept of conjugation of visual variables as proposed in the work by Roth (2017) allows to mix two variables and create visualisations that can be used to represent confidence or uncertainty of the same degree. The Figure 7.3 (a) and (b) explains how the manipulation of both size and degree of blur visually represents a confident and uncertain prediction (matched to the same degree) would look for two symbols - circle and arrow. One important observation that was made when applying conjugation of size and crispness for symbols was that when both variables are conjugated to the highest degree (x+2, y+2), the crispness was dominant over size with symbols appearing smaller in size that the original size variant (y+2, x). This influence was more prominent for arrow symbol considering its directional property. This has been a limitation of the conjugation approach used in our work. To overcome this current limitation, we have altered the size of symbols of higher degrees which is further explained in our Evaluation section.

H3 -Scenes: *Irrespective of the scene prediction complexity, all people walk similarly when seeing future paths in AR during motion conflicts.* This Hypothesis is formulated to see whether the results for the study can be generalised to all walking scenes and spaces, which is further tested separately for each of the complexities described above (Low, Medium and High).

7.1 Study Design

To test the effects of both categories of AR prediction (Simple and Informative) we test both simple and informative AR for Hypothesis H1. In *H1 test for Simple AR*, the study evaluated the difference in walking behaviours when there is no AR content against AR visualisations for future path for persons in a scene. For the *H1 test for Informative AR*, the study similarly evaluated the differences in walking seeing a confident prediction and

comparing them to uncertain predictions of scene users. The same test could also be used to infer behaviour differences for *H2 test*. The *H3 test* was completed by including scenes / environments of different complexity that are detailed in the next section.

Our mixed study for the work included both a within participant and between participant study to test the different hypothesis. The **within participant study** primarily included the *H1 test for Simple AR*. The **between participant study** comprised both the *H1 test for Informative AR* and *H2 test*. The complete study data was used to validate the third *H3 test* hypothesis. Hence in the complete study, participants of each Group tested one of the three condition - No AR, Simple AR prediction and a variant of Informative AR prediction (based on the Group) as shown in Figure 7.4.

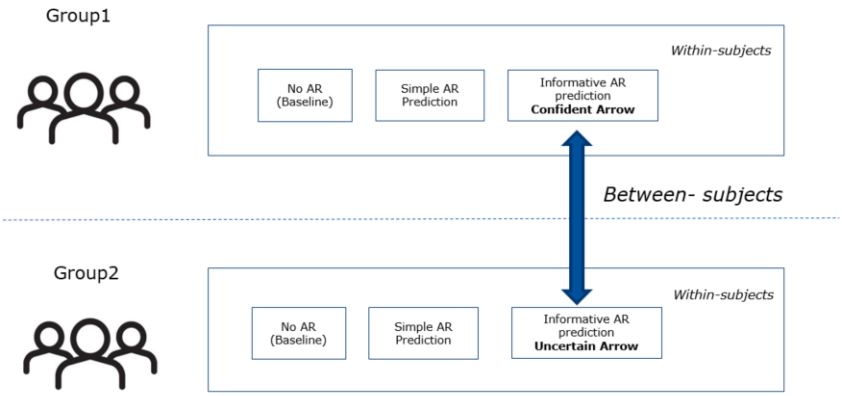
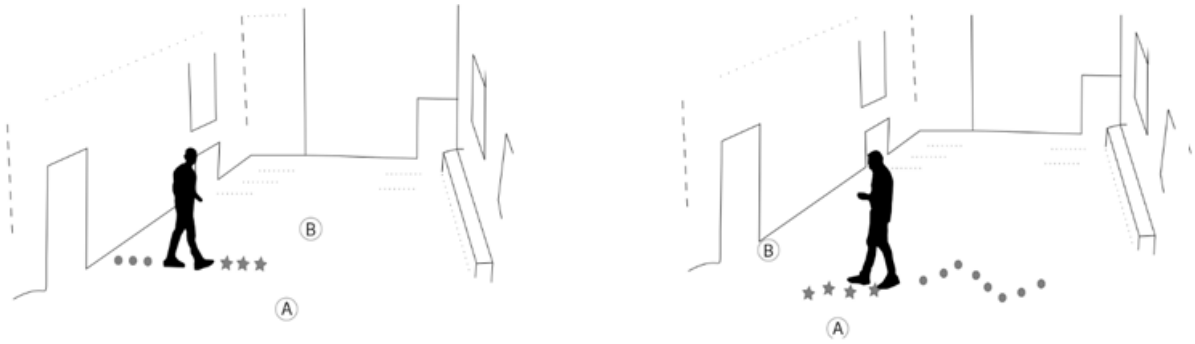


Figure 7.4: User Study design for future trajectory visualisation study.

7.2 Web Based Study

As the idea of testing the Hypothesis H1, H2 and H3 is not a requirement in real time, we designed the research as a web based study with post processed videos and AR content. For this, we firstly designed walking scenes based on different levels of scene & prediction complexity and created walking video sequences. We then include AR content to these videos to indicate future path for persons walking in the clip. Lastly we perform an online study were participants saw the created AR videos and made path choices to walk around the person in the clips that were then captured via sketches (virtual paths) on an image.

Scenes for the Web Based Study: The Figures 7.5, 7.6 and 7.7 show the representation of the scene motion, where the dots represent the path, which is shown to the ego-user, and the stars indicate the future path, which is shown by the different visualisations of simple and informative future. Considering the position of the person in the figures, the ego-user has to draw his/her own preferred future path between points A and B. These two endpoints, indicate the start and endpoint of the ego-user path. Traversing the path would potentially lead to a conflict with the person in the figure. The difficulty of predicting the future steps of the person varied for each of the captured scenes and were categorised as Low, Med and High. For illustration and better understanding, the scenes are labelled based on the categories with the subscript indicating the scene number.

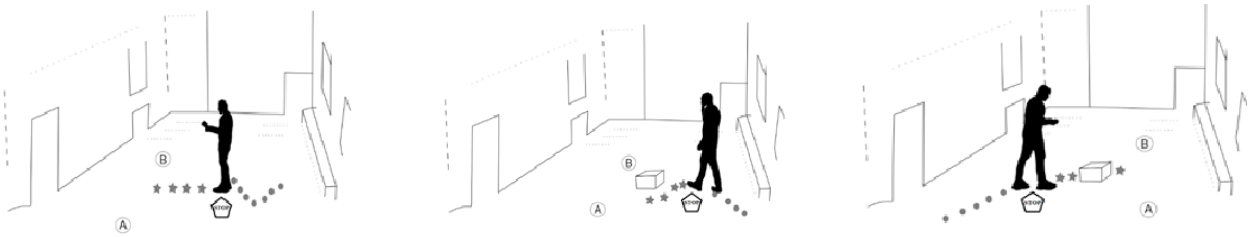


(a) AR prediction- continuation of path.

(b) AR prediction- continuation of path.

Figure 7.5: Low prediction complexity scenes Low_{01} and Low_{02} on the left and right respectively. The visualisations with the arrow showed the above mentioned motion prediction effects.

Low Complexity Scene - Low_{01} and Low_{02} Both the walking sequences captured for this complexity level represent motion along a predictable path. The single person in the video was seen to walk along a straight line path (Low_{01}) or in random paths characterised by zig-zag movement (Low_{02}) as in Figure 7.5. The scenes were characterised by lower difficulty in predicting future steps such that visually observing the motion of the past would be enough to estimate his/her future path.



(a) AR prediction- walk along straight line path.

(b) AR prediction- crossing from the front of the object.

(c) AR prediction- crossing from the back of the object.

Figure 7.6: The Medium complexity scenes Med_{01}, Med_{02} and Med_{03} from left to right. The visualisations with the arrow showed the above mentioned motion prediction effects.

Medium Complexity Scene - Med_{01}, Med_{02} and Med_{03} The three walking patterns for the medium complexity scenes represented a slightly higher level of difficulty in guessing where the person (shown in Figure 7.6) would continue walking. This was because the person stopped at the end of the walking sequence with his body posture only partially indicating where he might walk next. Visual observation of the past motion along with the body language (orientation of arms, face etc.,) was needed to make an estimate of the future path.

High Complexity Scene - $High_{01}$ and $High_{02}$ The high prediction complexity videos contained walking sequences (Figure 7.7) captured for a single person where visual observation alone might not have been enough to guess his future walking path. In High_{01}, the person is walking straight toward the wall and stops in front of it. Similarly, for High_{02}, the person

(a) AR prediction- walk towards left.

(b) AR prediction- walk towards right.

Figure 7.7: Higher prediction motion complexity scenes $High_{01}$ and $High_{02}$ for person with unpredictable change in walking path. The visualisations with the arrow showed the above mentioned motion prediction effects.

is initially walking toward the ego-user and then stops. The visual cues that are observed about both motion and body language might not be enough to make a safe future guess. In absence of such aids, extra information provided by the visualised future motion would help to guess whether the person would actually create a conflict with the AB path.

Furthermore, for all of the above stated scenes of complexity, the person in the video sequences resembled a distracted pedestrian performing some tasks and unaware of any potential collisions (Table 7.1) in future point of time.

Scenes	Interactions of the person in the scene
Low_{02}, Med_{02} and Med_{03}, $High_{01}$	Talking or Texting on smartphone
Low_{01}, $High_{02}$	Looking constantly or waving at the camera
Med_{01}	Reading a map

Table 7.1: Distraction for the person in the video sequences Low_{01} - $High_{02}$.

Visual Interface for Study To represent the future walking steps of the person in every timestamp of the video sequences, we use the arrow symbol and its 3D representation for AR visualisation. We choose a 3D arrow as the position and direction indicator of the future path and represent this as the **Simple Variant**. The **Confident Variant** was characterised by varying the visual variable length of the 3D arrow to emphasise higher confidence. This signified that the longer the 3D arrow, the more confident the predictions are. We further adopt the both the visual variable length and crispness to represent uncertainty. For this we vary the degree of blurriness of the boundaries of the 3D arrow which then represents the **Uncertainty Variant** in our study. Also, as the objective of the study was only to vary the attributes, a common color was used for the arrows throughout the work (Figure 7.8).

Using the Unity software, AR videos with the above mentioned arrows were created for the study. The animated arrows then represented the future positions of the person for every time stamp both in the case of simple AR prediction and Informative AR predictions (confident and uncertain respectively). Moreover raw videos which did not have predictions

Figure 7.8: Arrow for Simple(left), Confident(center) and Uncertain(right) representations.

of the future path were also used in the study. This was to used as a baseline for comparison during analysis.

Following the content creation, the videos where uploaded to two separate study groups websites. The first study group would complete the experiment containing a combination of no AR, simple AR prediction and variant1 of the Informative AR prediction (Baseline -Simple Arrow -Confident Arrow). The second link created for the other group contained no AR, simple and variant2 of the Informative AR prediction (Baseline -Simple Arrow - Uncertain Arrow). The video clips where also randomised to avoid any learning bias in the study. Each study link also contained a tutorial video at the start explaining the procedure of the study and a feedback questionnaire at the end following the study.

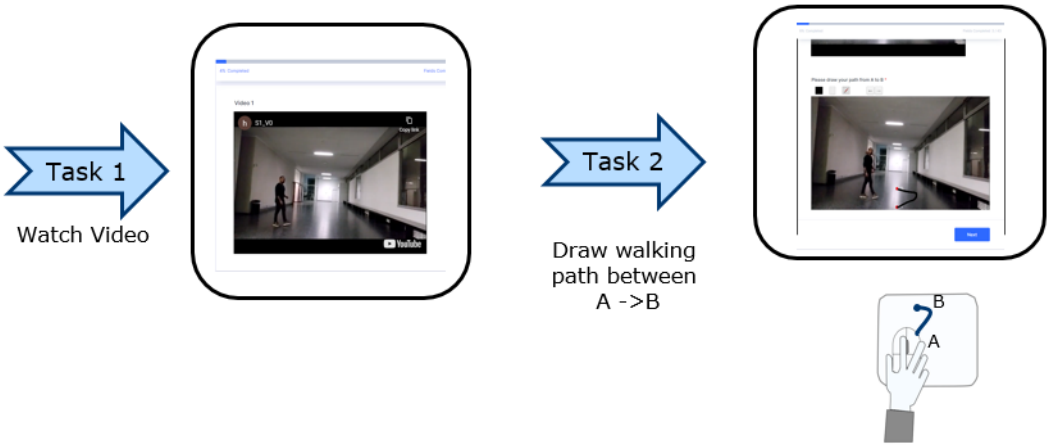


Figure 7.9: User study procedure for web based study.

Experiment Task The Figure 7.9 shows the procedure of the study that was conducted online. In the study, each of the participants first watch the created video sequence with or without AR (Task 1) and then draw paths on an image that followed the video (Task 2). The image shown in Task 2 corresponded to the last timestamp frame of the shown clip in Task 1. The participant in this tasks draw walking paths they would take for the video scene by estimating potential collisions with the person in the video. The resulting path drawn would then be a collision free walking path. To have uniform trip progression, participants where instructed to draw paths starting at point **A** that would then progress to point **B**. Both of these points **A** and **B** were marked on the drawing image as in illustrations in the Figure (7.5, 7.6 and 7.7).

7.3 Study and Data Extraction

Twenty-seven (n=27) participants each took part in the online study that was conduced in two groups. Of the total 54 (27 x 2) participants, study included candidates mostly from Germany and the rest of Europe.

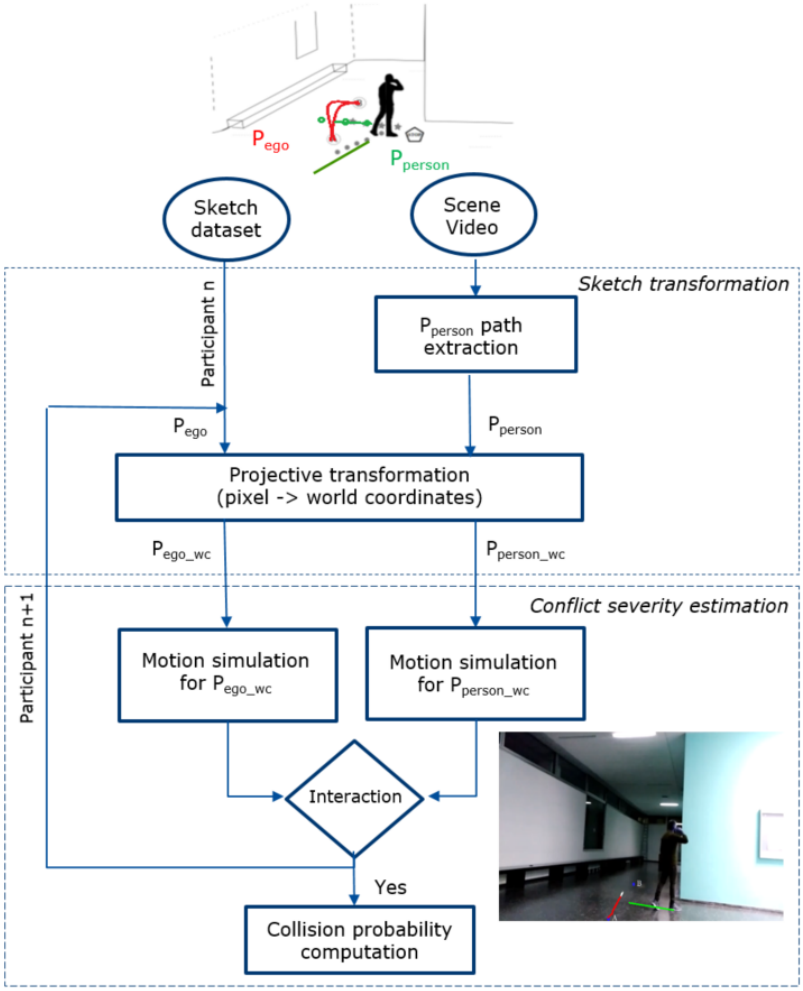


Figure 7.10: The participant sketch data extraction.

As the study examines walking path choices provided as drawings, the traced paths (in pixels) had to be transformed to motion trajectories. To understand how these virtual paths represent collision avoidance behaviour, we followed a two step data extraction procedure (Figure 7.10):

1. Sketch Transformation: Each of the sketches were converted from image coordinates to world coordinates (in meters) by a projective transformation [1] from opencv and ground control points measured in the scene. With this conversion, all the traced participant paths on images (P_{ego}) were transformed to *preferred ego-user walking trajectories* in world coordinates ($P_{ego_{wc}}$). We also extracted the walking paths of the person shown in the video clips (P_{person}) using Yolo v3 (Redmon and Farhadi 2018) image detection and represented the paths in projective transformed world coordinates ($P_{person_{wc}}$).

[1] *https : //docs.opencv.org/*

2. Conflict severity estimation: We apply motion animation to both the paths of the ego-user and the walking person in the video to estimate if there would be a walking conflict. For this, the transformed participant sketches ($P_{ego_{wc}}$) and the paths of the walking person ($P_{person_{wc}}$) were simulated using constant velocity. The time it took for their walking paths to cross each other was then estimated as time t_1. Using the corresponding time t_2 it would take for the other person to also pass the same conflict point, we captured conflict severity using PET (Post Encroachment Time) in seconds (Equation 7.1).

$$PET = t_2 - t_1 \tag{7.1}$$

Following the sketch transformation and simulation, all the sketches from the different participants were transformed to PET values in seconds. Sketches for one of the scene Med_{02} were excluded due to errors in accurately transforming the drawing paths.

7.4 Data Analysis

We perform our data analysis by separating scenes based on the different levels of complexity and then studying how the two groups (G1 and G2) reacted to the different variants of the AR predictions for each of the scenes. This was mainly because an initial comparison of the PET distribution for no AR visualisation (Figure 7.11) showed that in the absence of any future path clues, the people reacted diversely across the scenes. Three of the six scenes had PET values close to one for both groups while for the three, the values where less than one second. As 27 participants completed the study in each group, we had 27 responses for each of the "confident" and "uncertain" informative AR and 54 trajectories of preferred walking paths (27+27) for both the "baseline" and "simple" AR. This resulted in 54 and 27 PET values per scene to analyse user responses for "simple" and "informative" AR respectively. In most cases, the PET values had a range from 0 to 3 seconds.

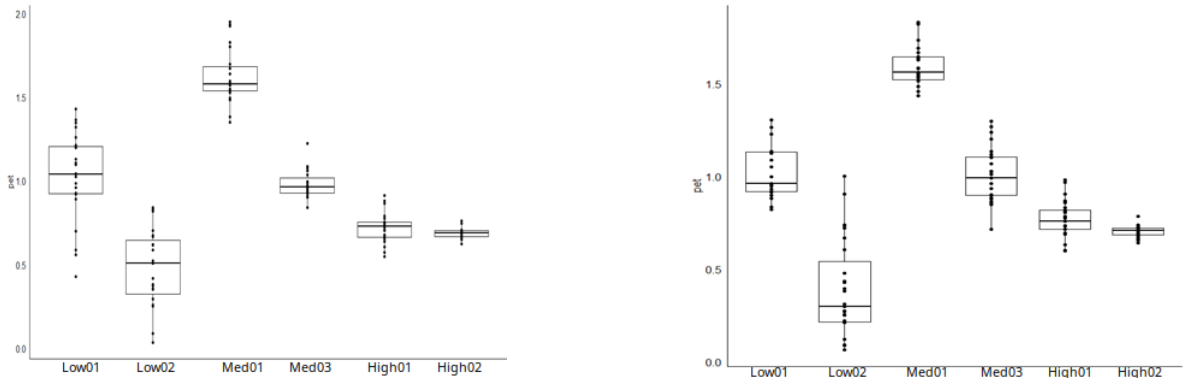


Figure 7.11: Figure on the left shows the box plot for Group 1 and compares it to Group 2 plot for the calculated PET values for the different scenes.

Crossing Order during Collision Avoidance: The walking behaviours from a collision avoidance prospective could be completely different when a person chooses to cross either in front or from behind a potentiaconflict encounter. **Front Crossing** (Figure 7.12) might result in an unsafe encounter as the agents are still facing each other and low PET values then

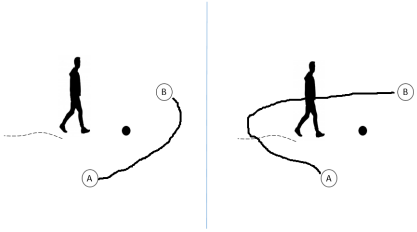


Figure 7.12: Illustration of classifying the collision avoidance based on either crossing front (left) or from the back of the person (right). The conflict point is indicated by the dark circle in the center.

would indicate that the threat is not completely avoided. On the other hand, **Back Crossing** could mean a safer encounter but with a discomfortable experience. One might brush past his/her shoulder when crossing the other person from the back. Crossing orders also indicate psychological traits (like arrogance or stubbornness) during conflict negotiations as studied in (Randhavane et al. 2019). Hence we further differentiate our sketch transformed PET data based on the crossing order. If the participant ($P_{ego_{wc}}$) decided to cross the person ($P_{person_{wc}}$) by walking in front or from the behind, we classify the resulting conflict interaction as *front crossing* or *back crossing* PET respectively.

Behaviour Shift Estimation: As the PET values provide information on how safely the participants encroached the person in the video, we modelled the complete experiment data as histograms (e.g., Figure 7.13 for scene Med_{01}) for further analysis.

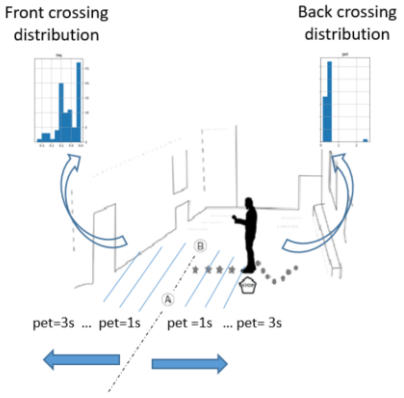


Figure 7.13: PET distribution as histograms for front crossing and back crossing behaviours.

The histograms have been used for comparing how the user's walking behavior changes when different visual stimuli are applied. While bin-to-bin distances quantify the similarity of histograms (Van Gemert et al. 2009, Ling and Soatto 2007, Marszałek et al. 2007), they do not represent the amount of correlation in the data. In our work, we were interested in understanding how PET histograms changed relatively for the same population for the different visualisations. Hence given two histograms, we approximate how similar they are by computing the time shift needed to move the bins of either histograms so as to achieve maximum correlation coefficient. As illustrated in Figure 7.14, after computing the bin-wise L1 distance between the two histograms, we apply cross correlation to estimate how the patterns represented by the difference in bins have shifted. Then the amount of PET shift that gives maximum cross correlation indicates whether the behaviour moved towards more

positive PET values or vice versa. A positive shift then indicates that people show more positive safety behaviour and a tendency to walk longer paths.

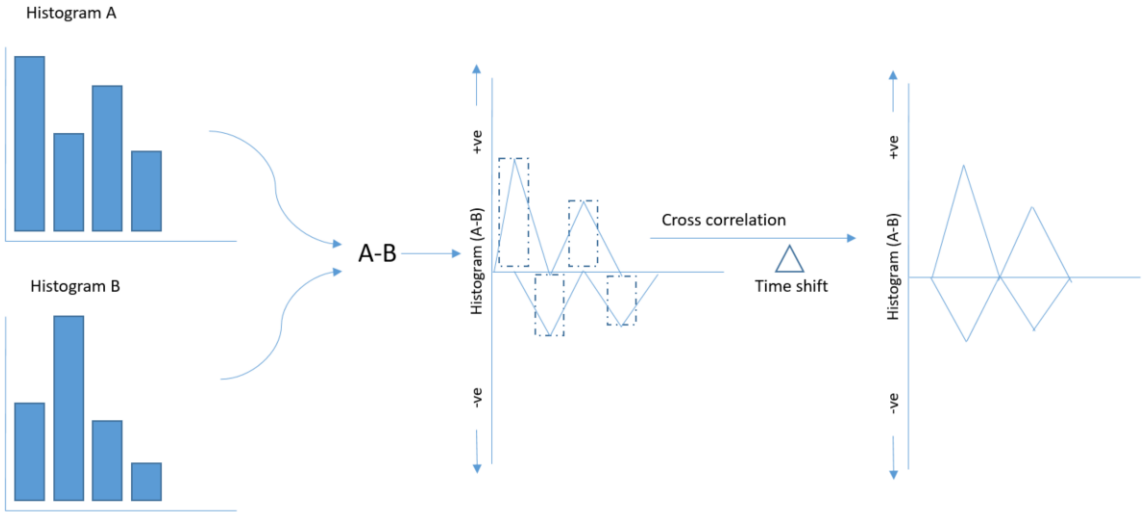


Figure 7.14: Histogram similarly computation using time shifts. A difference histogram is obtained by subtracting A and B. The set of positives and the corresponding set of negative values are correlated by the shifting of bins to either right(+ve) or left (-ve).

To apply this analysis, all PET values between 0-3 seconds were represented using histograms with a bin size of 0.25 sec. Furthermore, we separately analysed the PET distributions for both front and back crossing behaviours in each scene based on the visual variants. For the hypothesis H1, we performed analyses where the 54 PET values for "simple" AR visualisations are compared to the baseline (no AR) for the different scenes. Also, to investigate the effects of showing different variants (H1), we compared how people react to both "confident" and "uncertain" arrows by looking at the 27 PET values of both variants. Furthermore a PET density based analysis was completed to evaluate the second hypothesis for the two variants of Informative AR (H2). The results from the study as interpreted for the different scenes was used to test hypothesis H3.

7.5 Results

Walking Preferences during Collision Avoidance with and without "Simple" AR The computed shifts in PET behaviours when comparing "simple" AR to its baseline for both the unsafe front and the safer back crossings are shown in Figure 7.15. A positive shift indicates that the ego-user walked more safety-consciously as compared to the no AR condition and a negative shift indicates the tendency of the ego-user to walk closer to the person.

Front Crossing: For most of the scenes (except scene Low_{01}), the PET behaviors with "simple" AR showed positive shifts. This strongly indicates towards people interacting with a higher safety awareness (i.e., higher PET values), as compared to how they walked in the absence of any future visualisation.

Back Crossing: On the contrary, in nearly half of the scenes (Low_{02}, Med_{02}, Med_{03}), the interaction behaviour resulted in lower PET values as indicated by the negative bin shifts.

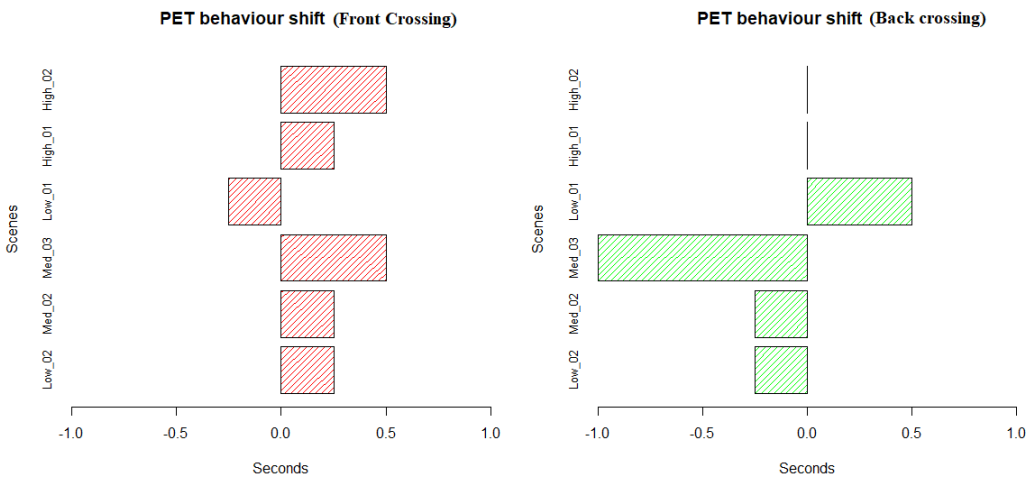


Figure 7.15: The time shift in front crossing (left) and back crossing (right) when comparing AR vs Baseline histograms for the different scenes.

For these three scenes persons decided to cross less safely (PET shift -0.25 for both Low_{02} and Med_{02}, -1.0 for Med_{03}) and potentially closer to the conflicting person even after the collision was avoided. Moreover for both Low_{02} and Med_{02} scenes, the shift was the same, indicating less safe behaviours in the presence of increasing complexity. For the back crossing case, high complexity scenes did not contribute towards significant shifts to indicate how the behaviours changed. This is primary attributed to fewer participants choosing to cross from behind for these scenes.

The primary observation from front crossing and back crossing PET trend highlights that people prefer to choose safer, but also longer paths when crossing from the front, and less safer shorter paths when crossing from the behind. Hence, we conclude that simple future predictions prompt persons to choose walking paths differently based on the crossing order, longer walks for unsafe and mostly shorter for the safer encounters around the conflict point.

Walking Preferences during Collision Avoidance for Different "Informative" AR To compare and interpret the walking preferences during a motion conflict for "informative" AR future visualisation, we evaluate how the PET values of 27 participants of Group1 (confident arrow) differed from Group2 (uncertain arrow). For this, we applied our histogram behaviour analysis where the shift for the "confident" histogram was estimated relative to the "uncertain" one. A positive shift in values then indicates that the ego-user decided to walk more safety consciously for the "confident" variant. The negative shift on the other hand, relates to safe walking for the "uncertain" variant of future arrow.

Front Crossing: As it can be interpreted from the PET shifts shown in Figure 7.16, the reactions show mixed behaviours for the different complexity scenes. For both low complexity scenes (Low_{01}, Low_{02}) and $High_{01}$ scene, the PET behaviours shifted by -0.5,-0.75 and -0.25 respectively. For all other higher prediction complexity scenes, the shift was positive (Med_{02} $\sim$+1.0, Med_{03} $\sim$+0.25 and High_{02} $\sim$+1.25). Our results for the analysis from safety and scene complexity prospective are two fold. People preferred to behave less safety-consciously (negative shift) for a confident arrow in most low complexity cases. As

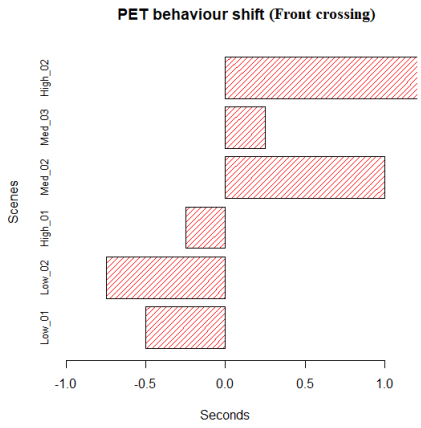


Figure 7.16: Time shifts in front crossing behaviour when "confident" visualisation is compared to "uncertain" future path visualisation.

the scene complexity increased, persons preferred to walk more safety-conscious (positive shift) for the confident AR. We hypothesise that the prediction complexity of the scene, the future path clues from the AR visualisation and the communicated certainty information played a role in the walking decisions made.

Due to the smaller sample size of the population (n=27) for the analysis and the resulting lower choices to cross from the back, statistical analysis was limited to front crossings.

Path Choices and Walking Preferences for the Different Visualisations To understand how the appearance of the future steps resulted in different walking and safety behaviours, we compared the front crossing PET distribution of each scene by approximating it using Gaussian density. The density estimation for back crossings were not considered primarily due to the lower sample counts.

Simple AR Visualisation: Figure 7.17 shows the density plots for the PET data for front crossing in the different scenes. The plot shows how the PET behaviour changed for both the baseline and "simple" AR conditions. For visual interpretation, the plot is flipped with PET =0sec on the right and PET>=3sec on the left of the x axis. The y axis represents the normalised density.

For low complexity scenes, it is evident that showing future predictions of the walking path lowered the inconsistency to decide on how far away to cross. For scenes Low_{01} and Low_{02} the PET responses were characterised by more distinct choices of the participants (as indicated by the peaks in the distributions) to either walk closer or further away from the conflicting person. This, we hypothesise could have been due to the extra aid provided by the future paths and larger confidence and trust in the visualisation.

But as the level of complexity of the scenes increased, the AR predictions where observed to be less influential to affect the choices made to the paths. This inference is concluded from the uni modal spread in PET distributions for both baseline and "simple" AR conditions for Med and High scenes. This indicates towards participants preferring no distinct crossing behaviours (as indicated by a bi-modal Gaussian spread of simple AR for Low complexity scenes) but making random choices to cross the person in the scene safety.

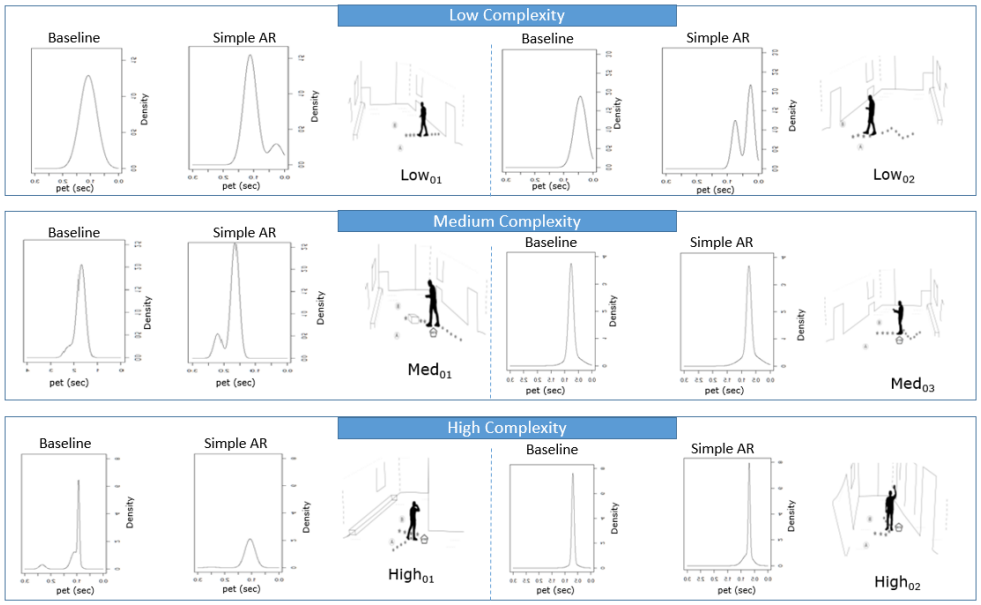


Figure 7.17: The PET based probability distribution for front crossing "Simple" AR Future path visualisation for low, medium and high complexity scenes. The y axis represents the normalised PET distribution and x axis represents PET values ranging from PET>=3 sec to PET=0 sec.

Informative AR Visualisation: A density plot based comparison (Figure 7.18) of the participant reactions to "confident" arrows did not show considerable behaviour differences to that of the "uncertain" arrow. However, if these plots are compared to the "simple" AR plots, informative AR collision avoidance behaviours were more distinct. This means that for a given scene with confident AR, if people choose to either walk closer or further away in the scene; this behaviour remained nearly the same for the uncertain future AR.

From the prospective of path choices made, even when safer paths would mean longer walks to avoid conflicts, our current study does not provide sufficient results to indicate whether a confident prediction resulted in longer or shorter walking paths. The visual representations of both informative variants need further evaluations to understand the interpret ability of the used visual metaphors (blur and length variation) to correctly communicate uncertainty to the user.

To summarise, the findings from the study is discussed from the perspective of how the results have supported or contradicted the stated hypothesis H1, H2 and H3. As for the safety hypothesis H1, if all front crossing behaviours are considered from the study, the simple AR visualisation prompted persons to walk away from the estimated conflict point resulting in longer walk paths. However when persons decided to give way and cross from the behind (nearly half the scenes of the study), people walked shorter paths while navigating from A to B and also crossing the conflict point more closely. Hence based on whether persons decided to cross from the front or from behind our results contradicted the hypothesis with only back crossings partially supporting it. Also when Informative AR was visualised, mixed reactions in terms of PET behaviours did not indicate towards people preferring shorter paths.

When the appearance hypothesis (H2) was tested using the density plot based visual comparisons (Figure 7.17 and Figure 7.18), it was clearly indicated that crossing behaviour with confident visualisation did not differ significantly from that of uncertain visualisations.

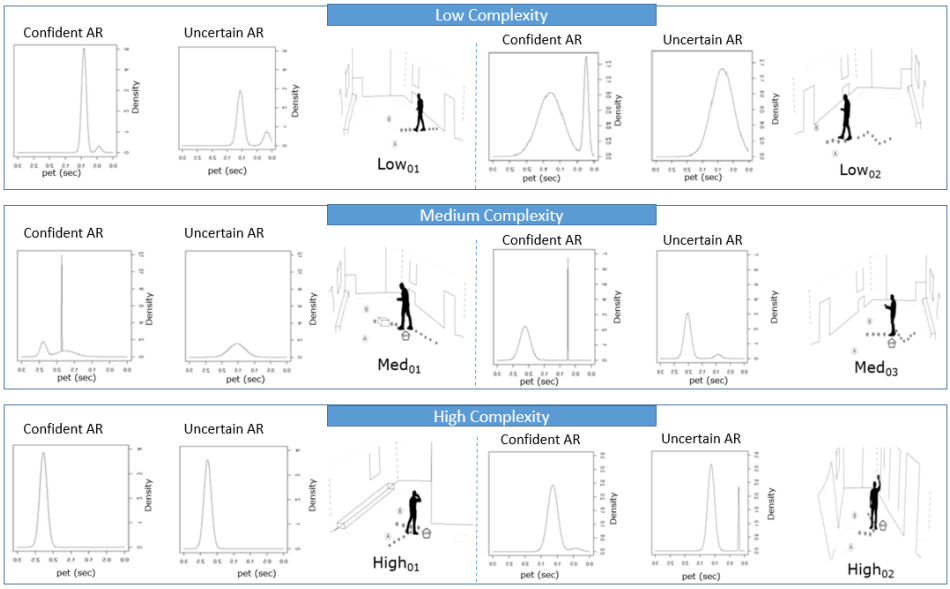


Figure 7.18: The PET based probability distribution for front crossing "Informative" AR Future path visualisation for low, medium and high complexity scenes. The y axis represents the normalised PET distribution and x axis represents PET values ranging from PET>=3 sec to PET=0 sec.

While this might partially contradict the second hypothesis, the study clearly showed informative AR prompting more distinct crossing behaviours over a simple AR visualisation. Especially for medium and lower complexity scenes, informative future paths resulted in bimodal distribution of PET data.

Furthermore the pet inference made for different scenes from Figure 7.11 and the density plot in Figure 7.17 strongly support hypothesis H3. As depicted by the box plot in Figure 7.11, each scene irrespective of the scene complexity had a different mean and variance estimate of how the sample population reached to collision avoidance. When AR future path was visualised , the PET Gaussian distribution in Figure 7.17 indicated similar dominant Gaussian peaks and data spread for the scenes. The above two observations strongly supported the hypothesis H3 that different PET distributions resulted for each scene and their reactions were similar even in the presence of AR future path.

8 Influences from Static and Dynamic AR Traffic Content

While the previous section examined how visualizing future paths and PET (Pedestrian Encounter Time) behaviors emerged from virtual content, this chapter shifts focus to augmented reality (AR)-enhanced walking—specifically from the perspective of pedestrian crossings. The proposed method in this thesis investigates how 3D-rendered virtual elements affect pedestrian path selection when pathways intersect and virtual content is present.

This chapter distinguishes between **static** and **dynamic** virtual content used to augment the Hololens user's environment. Static AR content typically recommends actions (e.g., indicating when to stop or cross), while dynamic AR content actively manipulates pedestrian behavior by encouraging a deviation from their original path.

The rationale for this distinction lies in the two types of navigational information they represent. Static content corresponds to fixed infrastructural elements such as traffic signals and lane markers—features that provide relatively stable control guidance. In contrast, dynamic content simulates transient or mobile entities like vehicles or cyclists, which prompt pedestrians to make adaptive, real-time decisions to avoid conflict or collision. An illustrative example of this adaptive behavior occurs during crossings. When a pedestrian perceives a moving virtual agent (e.g., a car) intersecting their path, they may modify their trajectory in both space and time—demonstrating dynamic avoidance behavior.

To evaluate the impact of these AR elements on pedestrian navigation, this chapter investigates their implications for safety in both walking and crossing contexts. Section 8.1 describes the methodology and experiments used to assess the effects of **static AR content**, focusing on virtual infrastructure visualization. Section 8.2 explores **dynamic AR content** and analyzes pedestrian collision avoidance behavior when encountering a virtually rendered crossing cyclist.

8.1 Influences from Virtual Infrastructure

Shared space designs aim to facilitate interaction among diverse road users by removing traditional traffic segregation and promoting informal communication. However, the creation of pedestrian-only zones—though often driven by safety concerns—can inadvertently isolate pedestrians from the integrated flow of urban movement. While urban designers frequently propose physical interventions (e.g., the addition of lanes or bollards) to enhance safety, there has been comparatively limited exploration of how virtual infrastructure, such as AR-based traffic signals or lanes, could contribute to safer interactions in mixed-traffic environments.

We hypothesize that the visualization of virtual traffic elements through devices like the Hololens can foster a heightened sense of control and safety among pedestrians. By clarifying right-of-way and reducing ambiguity over crossing priority, virtual infrastructure could help prevent confusion and enhance user confidence. Beyond safety, such AR elements may also enable new forms of priority negotiation in shared spaces.

For instance, at an intersection of pedestrian paths, a virtual traffic signal could dynamically resolve priority between a walking individual and a jogging pedestrian. Similarly, in a scenario involving a group of pedestrians and an approaching autonomous vehicle (AV) at a crossing, virtual signage could mediate the interaction—granting crossing priority to pedestrians based on context. Such interactions, mediated by augmented reality, may not only improve safety but also foster more socially acceptable and context-aware traffic behaviors.

Most existing research on virtual traffic infrastructure has primarily focused on creating virtual lanes to visually delineate pedestrian and vehicular zones. For example, an AR-based interface was prototyped by Hesenius et al. (2018), which presented separate walking paths for future automated traffic scenarios. This system dynamically adjusted the width of virtual walking lanes based on the pedestrian's walking speed, aiming to ensure safe navigation. However, their work was limited to a preliminary user study aimed at gathering initial feedback, without extensive evaluation in real-world traffic contexts. Similarly, Busch et al. (2018) conducted a technical feasibility study that explored the projection of virtual lanes to aid communication during priority negotiations. Their approach introduced virtual pedestrian crossings projected onto the ground to facilitate safe crossing in the presence of autonomous vehicles (AVs), focusing exclusively on crossing situations. To the best of the author's knowledge, no prior work has investigated the use of virtual traffic signals—visualized through AR—to influence pedestrian behavior and assess their safety impact in shared spaces. If integrated into daily traffic systems, such AR-based signals could guide pedestrians more safely in complex, mixed-traffic environments.

This research aims to fill that gap by exploring how AR-guided traffic controls affect pedestrian crossing dynamics, particularly in collision avoidance scenarios. A virtual traffic signal is designed in AR to recommend when a pedestrian should stop or proceed. The interface dynamically mediates right-of-way conflicts by displaying a green "go" or red "stop" signal when the Hololens user's path intersects with that of other agents in the environment. The experiment is designed to then investigate whether such AR mediating signals contribute to improving the time to estimation collisions. The system records their motion in real time, allowing for the extraction of Time to Collision (TTC) estimates based on walking trajectories. These trajectories are then analyzed to determine whether the AR-mediated signals contribute to improved collision avoidance behavior.

8.1.1 AR Virtual Signal Design

For this study, a tag-along virtual 3D AR signal, referred to as the *Stop and Go Interface (S&GI)*, was developed using Unity (Figure 8.1). In this context, a tag-along interface refers to a virtual object that remains persistently visible in the user's field of view and moves along with them, ensuring continuous visibility during navigation.

The S&GI was positioned at a fixed height above the ground and programmed to transition between two states: "Stop" (red) and "Go" (green). To control these transitions, an external control terminal was implemented. This terminal communicated wirelessly with the AR interface over a WiFi network and was manually operated during the experiment.

The design intention behind the S&GI was to simulate a conventional traffic signal within an augmented reality environment, prompting participants to proceed when the light was green and to halt when it turned red—mirroring the behavior expected in real-world traffic systems. However, to emulate an intelligent traffic management system that prioritizes

movement based on contextual factors, the control of the AR signal was manually administered by an external observer. This observer evaluated the scene in real time and adjusted the signal based on pre-established rules for priority during crossing interactions.

In this way, the AR interface approximated the decision-making processes of intelligent traffic control systems, such as those described in Chowdhury et al. (2018), which detect movement within a traffic scene and dynamically manage flow to optimize safety and efficiency.

(a) The red stop control of S&GI

(b) The green go control of S&GI

Figure 8.1: The S&GI interface commanding the person either stop (left) or go (right) while walking during the crossing experiment. The interface positioned at a fixed height and tagged to move along with the participant.

8.1.2 Experimental Design and Study

The objective of this study was to examine whether AR-based stop/go recommendations provided by the Stop and Go Interface (S&GI) could promote safer pedestrian crossing behavior. To test this hypothesis, the experimental walking path was intentionally designed to simulate potential collision scenarios between two individuals.

In each trial, a study participant walked along a straight trajectory, while a confederate—a person embedded in the experiment—intersected their path perpendicularly (Figure 8.2). This setup created a controlled conflict zone where crossing decisions had to be made. The S&GI interface, visible through the participant's Hololens, then issued real-time commands: either instructing the participant to stop or allowing them to proceed, depending on the predefined crossing priority for that trial.

At the beginning of each trial, the confederate initiated movement at an obtuse angle (approximately (180°)) relative to the participant's path and then gradually approached to intersect at nearly a right angle at a predefined cross-over point. This cross-over point represents the exact spatial location where the confederate's path intersects the participant's, and where a collision would occur if the participant failed to adjust their trajectory or speed in response.

This path was deliberately chosen to ensure that the confederate remained continuously visible to the participant throughout the interaction, thereby facilitating naturalistic and dynamic motion responses. Since collision avoidance behavior was anticipated to emerge in the vicinity of the intersection, a dedicated *interaction zone* was defined for the purposes of analysis. This zone was modeled as a circular area with a 2.5-meter radius, centered on the cross-over point—the spatial location where the confederate's path intersected the participant's trajectory. The interaction zone served as a key spatial boundary for subsequent evaluations of pedestrian behavior and safety outcomes during the experiment.

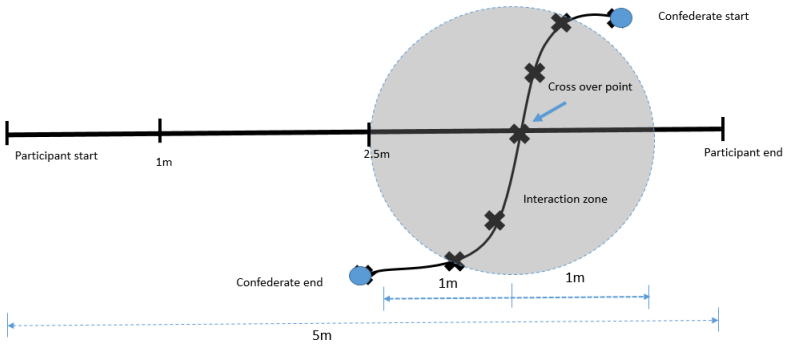


Figure 8.2: Participant start and stop points that are indicated on the floor along with the crossing path for the confederate. The confederate crossed the participant at the cross over point within the interaction zone.

The study employed a within-subjects 2 × 2 factorial design with the following independent variables: (1) confederate motion (crossing vs. non-crossing) and (2) AR guidance (no AR vs. signaled to stop). This design resulted in four experimental conditions:

1. Crossing confederate & no AR guidance
2. Crossing confederate & AR signaled to stop
3. Non-crossing confederate & no AR guidance
4. Non-crossing confederate & AR signaled to stop

Each participant experienced all four conditions in a randomized order to counterbalance order effects and reduce potential learning or fatigue biases.

The experiment took place in an indoor laboratory space measuring 5 meters in length and 3 meters in width. To capture motion data, an overhead camera was mounted at ceiling height, recording the participants' and confederate's foot positions at 30 frames per second. This setup enabled precise tracking of walking trajectories and subsequent analysis of behavioral responses across conditions.

A total of six participants (2 females, 4 males; mean age = 25.5 years) took part in the study. All participants reported normal mobility and had normal or corrected-to-normal vision. In addition, two trained confederates served as crossing agents, alternating roles across trials. Prior to the experiment, the confederates underwent training sessions to minimize variability in walking behavior and to ensure standardization of their motion patterns across all experimental conditions.

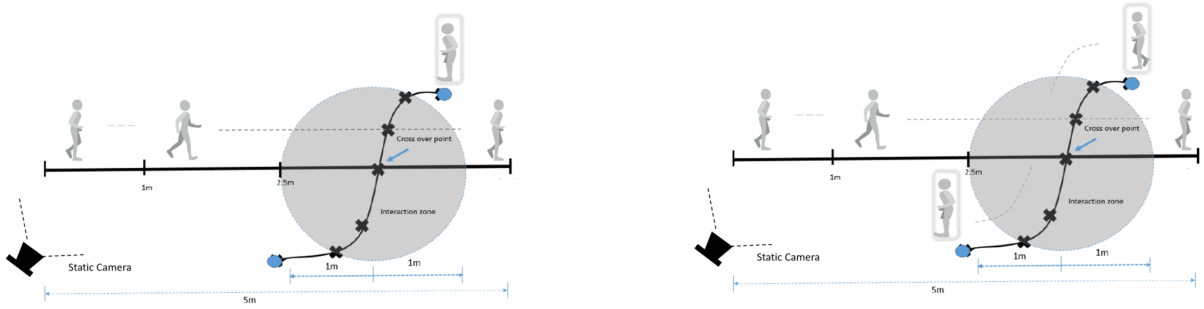
To maintain consistent walking speed, each confederate was equipped with a headset that played a metronome beat at 70 beats per minute. They were instructed to walk along the predefined intersecting path (as shown in Figure 8.2) and to synchronize each footstep with a metronome beat, resulting in a steady walking pace of one step per beat.

Throughout the experiment, confederates were instructed to maintain their trajectory regardless of the participant's position, including when a potential collision appeared imminent. Even if the participant entered the interaction zone, the confederate continued walking without reacting, ensuring that only the participant was responsible for executing a collision avoidance strategy.

Prior to the start of the experiment, participants were briefed on the study procedure and given an introduction to the HoloLens device as well as the Stop and Go Interface (S&GI). Detailed instructions were provided to ensure participants understood the purpose of the interface and how it might influence their actions during the task.

Participants were instructed to walk a total distance of 5 meters, following a straight path from a marked starting point to a designated endpoint. They were informed that, during some trials, another pedestrian (a confederate) might appear and potentially intersect their walking path without yielding. However, the specific trials in which this would occur were not disclosed, in order to maintain an element of uncertainty and elicit natural responses.

To ensure task comprehension and comfort with the AR system, each participant completed two practice trials. The main experiment commenced only after participants demonstrated a clear understanding of the task and system interaction. In the following section, we elaborate on the four experimental conditions that formed the different scenarios/trials of the experiment:



(a) No Interaction Motion Baseline.

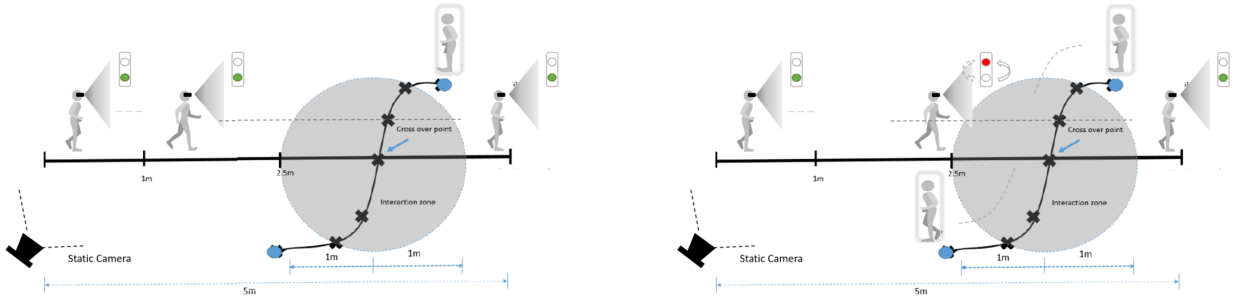
(b) No Interface Interaction Baseline.

Figure 8.3: The two no AR interface scenarios in the study where the participant exhibited natural behaviours.

Scenario 1: No interface - No interaction (No interaction Motion Baseline) This experiment scenario is characterised by the absence of both AR guidance and walking conflict interactions. The participant moved from the start to the end position without any S&GI control indicating to him when to stop walking. The confederate was instructed to remain stationary at his start position and did not interfere with the participant's path throughout the scenario (Figure 8.3a).

Scenario II - No Interface - Interaction (No Interface Interaction Baseline) This scenario was intended to capture the natural collision avoidance behaviour of the study subjects. The participant walked from the start point to the end without any visual control guidance while facing a conflict situation in between. In this scenario, the confederate was instructed to move and create a conflict by intending to block the participant's walking path as shown in Figure 8.3b. The study participant was expected to estimate the safety risks and take evasive manoeuvres.

Scenario III - Interface - No interaction (AR Interface Guided Motion) The participant in this scenario moved from the start to the end position following the guidance of the S&GI AR signals but with no conflict interactions with the confederate. The interface commanded the participant to start walking as it transitioned from red to green at the start of the trail. However, the interface remained in the green phase for the rest of the test. As no walking



(a) AR Interface Guided Motion

(b) AR Interface Guided Interaction

Figure 8.4: The two AR scenarios that would account for guidance effects of using virtual traffic controls

interaction was expected in this scenario, the confederate remained stationary at his start point and did not walk to cross paths (Figure 8.4a).

Scenario IV - Interface - Interaction (AR Interface Guided Interaction) In this final scenario, the participant moved from the start to the end position wearing the Hololens and guided by the S&GI AR interface. At the start of the trail, the external terminal triggered the transition of the traffic signal from "Red" light to "Green"; signalling the participant to start walking. Also simultaneously the confederate started to walk along his path to create a potential collision around the cross-over point. As both the participant and confederate entered the interaction zone, the AR interface instructed the participant to "stop" by transitioning from green to red. The wait time of the stop signal was based on the crossing of the confederate. Once the confederate walked past the cross-over point, the interface transitioned back to green signalling the participant to "go".

8.1.3 Data Analysis

To analyze participant and confederate movements during the experiment, overhead video recordings were processed using a computer vision pipeline. Person detection was performed on the video data using the YOLO (You Only Look Once) object detection algorithm (Redmon et al. 2016), which provided bounding boxes around individuals in the scene. These detections were then passed to DeepSORT (Wojke et al. 2017) for multi-person tracking and generating trajectory data for both the participant and confederate.

To convert the tracked pixel coordinates into real-world positions, a projective transformation was applied using OpenCV's geometric transformation functions[1]. This allowed mapping of the 2D camera view to the physical dimensions of the experimental space. To enhance the quality of the trajectory data, a median filter was applied during post-processing to suppress noise and remove outlier detections. Finally, the trajectory data was downsampled to 5 frames per second (fps).

To evaluate how motion dynamics and collision avoidance behaviors varied across experimental conditions, a trajectory-based safety analysis was conducted. Specifically, we extracted the walking speed of participants from their tracked trajectories and computed the Time to Collision (TTC) for each trial involving an interaction with the confederate.

[1] `https://docs.opencv.org/3.4/da/d6e/tutorial_py_geometric_transformations.html`

$$TTC = \frac{d}{V_f - V_l} \tag{8.1}$$

For the above equation, the instantaneous tangential velocity of the person that represented speed for each time stamp was then computed as

$$V(t) = \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}$$

where x and y denote the position of the foot of the person when representing motion in 2D world space.

The TTC was calculated using Equation 8.1, which estimates the time remaining before a potential collision occurs, based on the relative velocity and distance between the participant and the confederate. This metric then served as a key indicator of proactive avoidance behavior reflecting the effectiveness of the AR guidance in promoting safer pedestrian interactions.

In Equation 8.1, d is the distance between the participant and confederate and V_f and V_l correspond to the speed of the participant between consecutive frames proceeding a collision. While TTC has been used to represent the collision behaviours in the study, we also use the spatial distance d *or* GAP from the above equation to observe how the spatial separation between crossing persons differed for the different conditions of the study. Hence TTC, GAP, mean walking speeds and computed statistical values of standard deviations(SD) were used to make inferences in the work.

8.1.4 Results and Discussion

Figure 8.5 presents the extracted motion trajectories (shown in green and purple) for all participants in Scenarios II and IV, which involved encounters with a crossing confederate. As depicted in Figure 8.5(a), when no AR guidance was provided, participants exhibited varying collision avoidance strategies (track deviation or speed variation), with some deviating from their intended straight-line path to prevent serious conflict. The behaviors reflect spontaneous, self-initiated avoidance in the absence of external instruction.

In contrast, Figure 8.5(b) illustrates participant behavior under AR-guided priority cues. In these trials, the Stop and Go Interface (S&GI) provided explicit instructions to stop or proceed. It can be noted that participants consistently complied with the AR signals—stopping when instructed. This effectively helped to prevent collisions at the cross-over point. The results suggest that AR-based control guidance was effective in modulating pedestrian behavior and improving safety outcomes during path-crossing interactions.

The Time to Collision (TTC) and GAP distance which respectively characterize the temporal and spatial aspects of pedestrian-confederate interactions showed notable differences between AR-guided and non-AR conditions. As shown in Table 8.1, participants demonstrated significantly higher TTC values and greater stopping distances (GAP) when guided by the Stop and Go Interface (S&GI), compared to the baseline (no inference) condition.

In the No S&G Inference condition, GAP values varied across participants, indicating individual differences in perceived safety margins during conflict resolution. For example,

Figure 8.5: Interacting pedestrian trajectories in Scenario II (No Interface Interaction Baseline) on the left and Scenario IV (AR Interface Guided Interaction) on the right. The green dots represent the foot position of the participant, while the purple represents the trajectories walked.

participants P1, P5, and P6 maintained relatively consistent safety distances, suggesting personal strategies or habits in approaching crossing scenarios.

However, under AR-mediated priority negotiation via S&GI, GAP values were found to increase by at least 25%, highlighting the effectiveness of the S&GI in promoting more cautious and structured interactions. These findings suggest that AR-based signaling not only helps in avoiding collisions but also encourages participants to maintain safer distances, enhancing both spatial awareness and reaction timing.

Participant	Scenario	GAP (cm)	TTC (sec)
P1	No Interface	77.05	2.5
	AR Interface	106.01	3.5
P2	No Interface	111.01	3.7
	AR Interface	132.6	4.42
P3	No Interface	59	0.65
	AR Interface	104	3.4
P4	No Interface	32	0.36
	AR Interface	67	2.23
P5	No Interface	70.29	0.39
	AR Interface	95.18	3.17
P6	No Interface	72.11	0.48
	AR Interface	100.12	3.37

Table 8.1: GAP and TTC responses of participants for Scenario II (No Interface Interaction Baseline) and Scenario IV (AR Interface Guided Interaction).

To better understand how the crossing dynamics of the persons changed in the AR experiment, we studied the speed profiles of the different participants individually. This was then used to make interpretations of applying AR traffic signals to walking speed control and collision avoidance while crossing. The following research questions in this regard helps to address gaps to AR signal control and pedestrian crossings.

Q1: How does the normal walking behaviour of the participant change when controlled by AR mediation motion influence?

Table 8.2 compares walking speed data from Scenario I and Scenario III, both of which involves uninterrupted, straight-line walking without any interaction with a crossing confeder-

ate. The mean walking speeds and their associated standard deviations offer a quantitative measure for assessing whether participants modified their walking behavior in response to the presence of AR elements. Specifically, a reduction in average speed or an increase in speed variability may indicate a more cautious walking pattern, potentially reflecting elevated cognitive load or uncertainty introduced by the AR interface. In contrast, higher and stable speeds may suggest that participants were minimally influenced by the AR content. They were able to maintain normal attentional states during navigation. As such, these speed profiles provide valuable insight into the subtle behavioral impacts of AR systems, even under conditions without explicit pedestrian interaction or conflict.

	Scenario I		Scenario III	
	Mean Speed (m/s)	SD	Mean Speed (m/s)	SD
P1	0.79	0.46	0.53	0.277
P2	0.85	0.56	0.50	0.37
P3	0.60	0.30	0.43	0.21
P4	0.57	0.23	0.62	0.18
P5	0.53	0.37	0.55	0.25
P6	0.56	0.23	0.47	0.202

Table 8.2: Walking speed variations for different participants comparing the two scenarios where AR was present but did not control crossing behavior.

When comparing the **No Interaction Motion Baseline** with the **AR Interface Guided Motion**, it was observed that most participants exhibited slower walking speeds under AR-guided conditions. This suggests that participants generally modulated their motion dynamics, as evidenced by both reduced mean speeds and lower standard deviations during AR-assisted walking.

These findings indicate a shift toward more controlled and deliberate movement, which may reflect increased cognitive engagement with the AR interface. Such behavior aligns with previous research in AR navigation contexts, which has shown that users often slow down when required to continuously process and respond to virtual guidance cues (Tang and Zhou 2020). The observed reduction in speed may therefore be attributed to the elevated attentional demands imposed by the Stop and Go Interface (S&GI), reinforcing the idea that AR-mediated navigation affects not only decision-making but also baseline locomotion behavior.

Q2: How did the walking behaviors change from a collision avoidance perspective when being continuously guided with AR

As collision avoidance is a complex phenomenon that has been extensively studied, understanding the AR effects to avoidance required detailed analyses of strategies of how the person prevent an accident. Common counter-strategies in real-world walking include diverting from the obstacle, slowing down, or entirely avoiding the path of a potential conflict.

Pedestrian motion modelling via simulations (Johora 2022) too has accounted for different strategies to obstacle presence and their avoidance effect. In such simulating models, collision avoidance is modelled as a social force that would repel motion with the strength of repulsion propositional to the distance of the obstacle from the walking person. Taking into account the motion responses during a conflict and considering AR influences; we looked at the reactions of the participants to (a) the attention given to the **presence** of the nearby

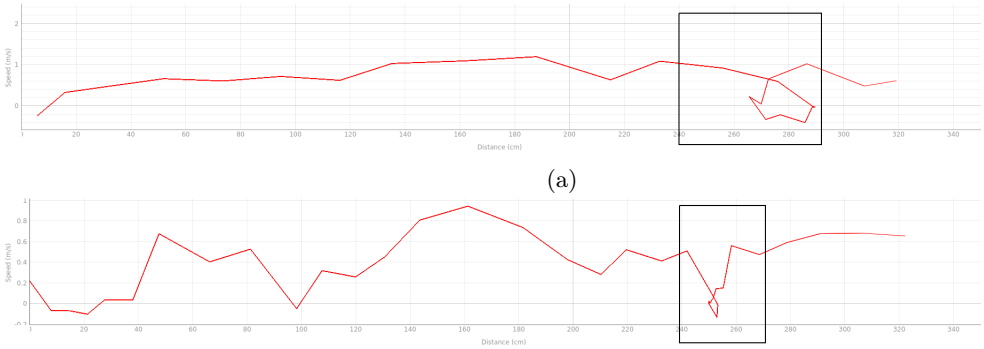
confederate when being controlled, (b) their **braking** behaviours when encountering crossings persons and (c) the resulting changes to **collision avoidance** based on path or speed adjustments.

Presence: To assess the effects of AR-mediated control on walking behavior, we compared participant motion in the No Interaction Baseline (Scenario I) with No Interaction under AR guidance (Scenario III). This comparison was used to evaluate how the mere presence of a potential conflict agent (the confederate) influenced pedestrian motion, and how this influence changed when AR cues were introduced.

Analysis of walking speeds revealed that two of the six participants demonstrated cautious behavior when approaching the cross-over point in the baseline condition—likely due to the anticipation that the confederate might initiate a crossing at any moment. This suggests a proactive adjustment in speed, influenced by environmental uncertainty and perceived collision risk. However, when the same scenario was repeated with AR signaling the participant to begin walking, this anticipatory behavior was not observed. In the AR-guided condition, all participants maintained consistent walking speeds, showing no slowdown in response to the presence of others in the scene.

This outcome highlights the strong directive power of AR instructions, indicating that virtual cues can override natural social and spatial heuristics typically used in pedestrian navigation. The AR interface effectively suppressed the cautious tendencies that participants might otherwise employ in uncertain, potentially conflicting environment suggesting a shift in behavioral authority from real-world cues to digital guidance systems.

Braking: To investigate changes in braking behavior as a means of collision avoidance, we conducted a visual analysis of the walking speed profiles of participants who exhibited braking responses during interactions with a crossing confederate.



(a)

(b)

Figure 8.6: The Figure (a) above shows the natural response of the participant wherein P4 stopped abruptly for the reaction window while Figure (b) shows the same participant reacting to AR interface stop more smoothly and quickly to avoid the same collision.

Our initial findings suggest that braking responses were more abrupt when participants were guided by the AR interface during crossing scenarios. Figure 8.6 illustrates the speed profile of Participant P4, whose behavior was closely examined within a defined speed reduction window (highlighted in black).

In the non-AR condition, the participant demonstrated a gradual deceleration, indicative of a cautious approach. Notably, the participant also exhibited a brief backward motion, interpreted as an additional maneuver to increase separation distance and mitigate collision risk. This behavior reflects a more deliberative and self-initiated collision avoidance strategy, likely driven by personal safety heuristics and real-time visual assessment. Conversely, in the AR-guided condition, the same participant responded more rapidly and decisively to the virtual stop command, with a sharper deceleration and minimal backward movement. This suggests that the presence of AR signaling prompted a more immediate and confident braking response, likely because the perceived responsibility for decision-making shifted from the individual to the system.

These observations support the inference that AR mediation can compress the decision-making timeline, leading to more abrupt but controlled braking behaviors, particularly in otherwise cautious pedestrians. Additionally, the reduced backward displacement under AR guidance implies greater reliance on the virtual interface for conflict resolution, potentially reducing the cognitive burden associated with evaluating dynamic obstacles in real time.

Collision Avoidance Strategies: While most of the participants preferred to brake giving the right of way to the crossing confederate, two participants (P4 and P5) reacted with both path and speed adjustments for No AR Interface Interaction Baseline (Scenario II).

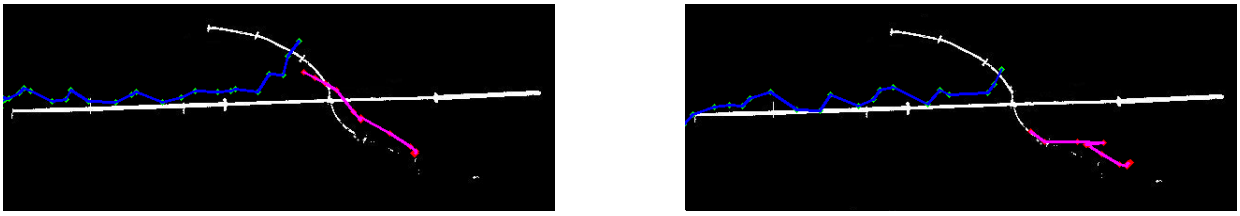


Figure 8.7: The left shows participant P4 exhibits path adjustment as the collision avoidance strategy while encountering the confederate, while P5 exhibits a combination of both path and speed adjustment to counter the crossing confederate.

Figure 8.7 shows the avoidance behaviours of these two participants confronting the crossing confederate (shown in pink). P4 applied a path adjustment along with upper body movement that was visually observed from the video clips. The P5 participant, on the other hand, applied a combination of both speed and path adjustments to avoid collision. In general, all the participants (P 1-5), when guided using AR strictly followed the AR interface, giving up participants specific avoidance strategies to avert the collision. This could indicate towards trust in the AR-mediated virtual crossings and the attention given to the presented information.

To summarise, this section detailed on how our exploratory study investigated the influence of AR virtual traffic to mediate pedestrian crossings. The results of the study indicated towards a virtual traffic interface successfully reducing the severity of motion conflict, affecting the crossing dynamics and also collision avoidance strategies while sharing space and moving with others.

8.2 Influencing Behaviour by Virtual Cyclists Visualisation

Moving objects in a traffic scene could sometimes be partially occluded due to non-light-of-sight visibility. Then a hazardous situation could arise when a person decides to walk or cross even in the absence of complete scene information. For example, when moving behind parked cars or walking past an intersection that is partially visible (due to layout or bus stop nearby), special care needs to be taken to avoid bumping into a crossing cyclist or other fellow pedestrians. Visual mediums play an important role in communicating scene information during occlusions. A network-based see-through approach was presented in (Olaverri-Monreal et al. 2010) where visuals of the road that were occluded due to visibility blocking trucks ahead helped drivers make safer overtaking decisions. The overtaking vehicle then had a better perspective of the road ahead which was possible due to video streaming of the scene and inter-vehicle communication. Extending such a see-through approach to AR, Rameau et al. (2016) applied sensor-based perception where stereo vision and localisation were then applied to create synthetic images that allowed drivers to see through occluding vehicles. A similar approach was then extended in (Maruta et al. 2021) allowing pedestrians wearing Hololens to see blocked objects with V2X communication. All of the above-mentioned approaches focused on directly communicating visual information of occluded objects and improved driving and safer traffic behaviour.

If Mixed Reality can be used to represent traffic participants (like a cyclist), then such virtual presentation methods could be useful in situations with incomplete visual information. For example, if a parked vehicle occludes an approaching cyclist, then a person crossing could walk more cautiously if s/he sees a virtual cyclist approaching him using his AR glasses. However for such an approach to be effective in influencing the motion path of the crossing person, the virtual cyclist should behave similarly to a real cyclist. This means that if a person steps into the path of a virtual cyclist, it should also not bump into the person and choose a collision-free path. Also, the visual appearance of a virtual cycle should prompt safer crossing reactions rather than persons avoiding its presence completely. Hence to prove the influences of mixed reality cyclists on walking, the work in this section focuses on first implementing cyclists as intelligent agents with collision avoidance behaviours and then testing whether their presence could visually influence walking. Section 8.2.1 describes the framework to demonstrate the interaction between a virtual cyclist and a real person blocking its path, Section 8.2.2 introduces the method proposed to estimate the level of influence AR content has when walking with virtual traffic agents in a real-world experiment. The methods described in this chapter are also presented in (Kamalasanan et al. 2022b) and (Kamalasanan et al. 2023).

8.2.1 Mixed Reality Agent Framework for Cyclist Interactions

The novel framework introduced in this section prototypes a mixed-reality (MR) agent system that enables real pedestrians to interact dynamically with virtual two-wheeler cyclists agents. The setup for the framework setup employs a dual-subsystem consisting of a primary and a secondary interface, as illustrated in Figure 8.8.

The primary interface, implemented via an AR headset, allows participants to visually perceive and interact with a virtual cyclist while moving through the real-world environment wearing the headset. Simultaneously, a secondary interface—a desktop-based computer station captures data related to the participant's movement and behavior during the interaction. Both interfaces are synchronized in real time over a shared network, ensuring

seamless communication between the AR system and the data-recording setup. While the primary interface facilitates immersive interaction with the virtual cyclist, the secondary interface serves as the observation and logging system, recording parameters such as device trajectory and timing of responses. By analyzing the recorded data, this framework enables a quantitative study of pedestrian–cyclist interactions, including conflict scenarios and avoidance behaviors, in a controlled yet ecologically valid MR environment.

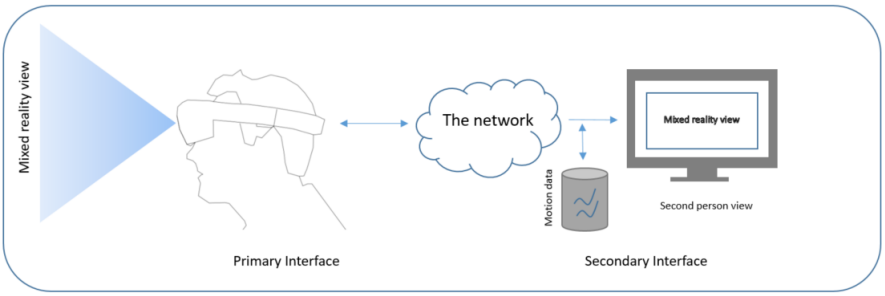


Figure 8.8: The primary and secondary interface for Mixed Reality Agent Framework.

Given our goal of investigating pedestrian collision avoidance through controlled experimentation, we developed a scene-centric framework. This choice is rooted in the understanding that pedestrian behavior is strongly influenced by the spatial configuration and semantics of the walking environment. As such, our system begins with the creation of a digital twin of the experiment site using MR development tools. This digital environment serves as the foundation for introducing virtual cyclist agents, which operate as autonomous, context-aware characters capable of responding to pedestrian motion. A digital twin of the experiment site is first created using mixed reality software and cyclists agents are then introduced in it as intelligent characters. Participants would then interact and test crossing the path of the virtual cyclist in the experimental scene. Such a *test scene* specific implementation was a prerequisite to achieving intelligent behaviours of our novel cyclist agents. Figure 8.9 depicts the three main steps in realising our cyclist agent framework:

Figure 8.9: The workflow proposed framework that includes modeling- the creation of a test site,MR Simulation- interactions with virtual content and the data capture from the interaction experiment using the secondary interface.

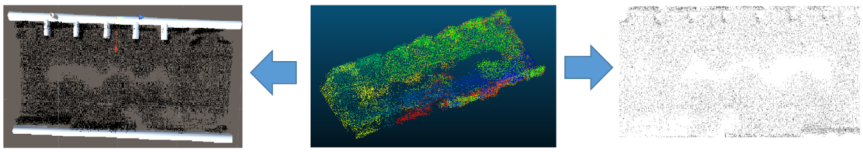


Figure 8.10: The capture point cloud of the test site is post-processed to model a 3D Unity mesh model (left) and further transformed and pixelated to create a 2D map (right) of the test site.

Modelling: The primary step of the framework focuses on capturing the experiment site either from a 3D or 2D perspective for virtual world creation and rendering. The created virtual replica of the experiment scene is used for virtual content alignment using QR codes in the primary Hololens interface. Furthermore the model in 2D prospective is also used in the secondary interface (Figure 8.10). The following details the modelling procedure :

- 3D Modelling: To generate a 3D representation of the indoor or outdoor experimental scene, a point cloud of the test environment was captured using the HoloLens 2 device. Specifically, we utilized the HoloLens 2 Research Mode (Ungureanu et al. 2020), which provides access to raw sensor streams, enabling the creation of a colorized point cloud that accurately reflects the spatial structure and appearance of the environment. Following acquisition, the raw point cloud data underwent post-processing to eliminate noise and ensure geometric consistency. The cleaned dataset was then imported into Unity[2]. Using the point cloud as a visual and spatial reference, a 3D mesh model of the scene was manually reconstructed within Unity. This reconstruction relied on Unity's primitive objects and GameObject system[3], allowing for a customizable and lightweight digital twin of the physical environment suitable for real-time rendering and interaction in MR.
- 2D Modelling: The raw 3D point cloud captured during the scene modeling step was subsequently transformed into a Bird's Eye View (BEV) map to obtain a 2D top-down representation of the test environment. This transformation enables a simplified yet spatially consistent layout of the scene for further analysis and agent navigation. To achieve this, the point cloud was first downsampled and voxelized to reduce computational complexity while preserving essential geometric features. Following this, the 3D coordinates were projected onto a 2D plane—typically the ground plane—by mapping the x and z spatial dimensions to pixel coordinates, with the vertical y-axis (height) discarded. This pixelation step effectively converts the volumetric data into a 2D grid map, enabling top-down visualization and facilitating subsequent tasks such as path planning, agent localization, and collision detection within the mixed-reality framework.

Mixed Reality Simulation: The simulation of motion for the virtual cyclist to enable real-time collision avoidance is a core component of this stage. This is accomplished by integrating multiple software modules encompassing position tracking, agent-based modeling, and visualization (as illustrated in Figure 8.11) within the broader mixed reality (MR) simulation pipeline.

[2] https://unity.com/
[3] https://docs.unity3d.com/Manual/GameObjects.html

The MR simulation operates under a client-server architecture, wherein the HoloLens device functions as the client responsible for rendering the visual scene, while a centralized server handles the real-time movement logic of the cyclist agent and facilitates the communication required for synchronized viewing. During the simulation, the primary HoloLens user—immersed in the MR environment continuously transmits their positional data to the server. This data acts as a dynamic input to the agent-based path planning algorithm, which governs the motion of the virtual cyclist. The cyclist agent, managed on the server, updates its trajectory in real-time based on the user's movement and employs a collision avoidance strategy informed by proximity-based interaction rules. The computed positions of the cyclist are subsequently streamed to the client and rendered in the HoloLens as 3D animated avatars, enabling the user to experience realistic motion encounters and conflict scenarios. This framework allows for context-sensitive interaction, simulating intelligent and reactive behavior of cyclists in the shared MR space.

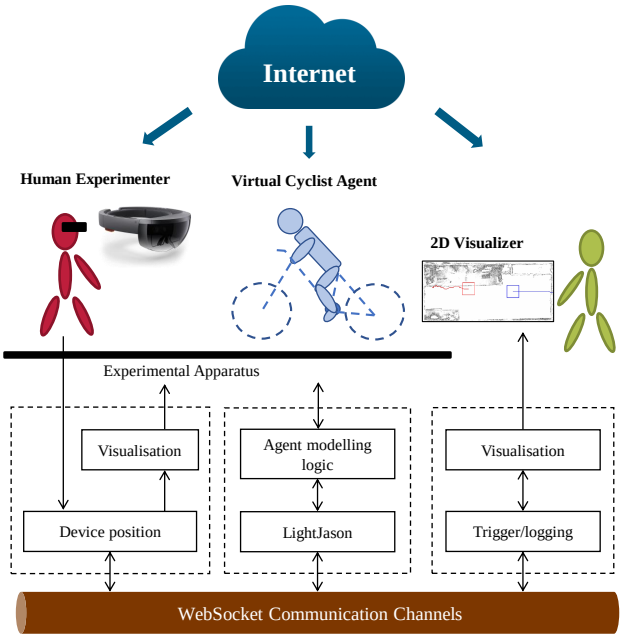


Figure 8.11: Different components of the real and virtual world (human experimenter, virtual cyclist and 2D visualizer) interconnected using web sockets in the mixed reality simulation.

Data collection: The motion interactions occurring between the primary HoloLens user and the virtual cyclist are monitored and recorded through a secondary interface. As illustrated in Figure 8.9, this secondary interface includes a 2D visualizer integrated with a 2D map of the test environment. Leveraging real-time positional updates from both the HoloLens user and the simulated cyclist agent, the visualizer provides a live display of motion trajectories and interaction events.

This utility interface serves a dual purpose: it facilitates real-time monitoring of user-agent interactions and enables post-experiment analysis by recording motion data for later replay and evaluation. Such a capability allows researchers to examine pedestrian-agent dynamics in detail, supporting both quantitative assessments and qualitative reviews of behavior during mixed reality simulations.

Prototype implementation: To prove that our framework can successfully capture the interaction between a real person interacting with a virtual cyclist, we implemented our mixed-reality apparatus using MRTK, web sockets and Unity. To perform a microscopic simulation between the real person and the virtual cycle, we chose LightJason (Aschermann et al. 2016) as the agent framework and implemented the server logic using Java. The cyclist collision avoidance was then achieved based on real-time position updates from the Hololens and the LightJason backbone. All simulation movements of the software agent were based on a 2D grid-based occupancy of the Hololens in the scene. This means that whenever a person occupied one of the grids, then the cyclist found a collision-free path avoiding the occupied grid. The agent description of the cyclist was written in AgentSpeak (L++) language and the Stomp protocol was used to publish and subscribe the different messages between the different components.

To interact with the virtual cyclist, the primary interface user would first launch the Mixed reality application on the Hololens device and scan the QR codes placed in the experimental scene. This would align the Unity coordinates system with world coordinates using the 3D model of the scene as in (Hübner et al. 2018). Once the real-world scene is aligned with the virtual world, the virtual cyclist would start motion from its origin moving towards its destination via predefined way-points that were set in the experiment. The human experimenter viewing the cyclist would walk and block the path of the virtual cyclist during the experiment. Based on the position updates from the Hololens, the programmed agent-based simulation would route the cycle avoiding the occupied positions along its motion paths. The 2D visualiser that is part of the secondary interface would receive all the motion updates of the experimental apparatus in real-time. The interface also contained a GUI to view the interactions and save them for further analysis.

Figure 8.12: A Mixed reality view from the Hololens of a virtual cyclist moving and interacting with the AR headset user.

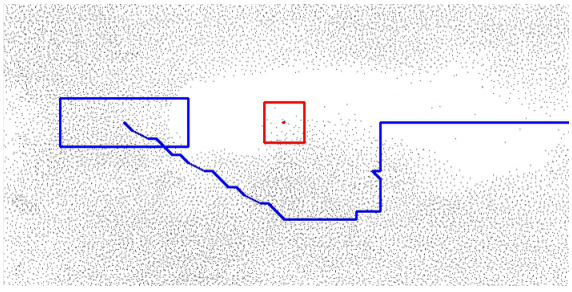


Figure 8.13: The virtual cyclist (blue) forced to take a detour as a static Hololens experimenter (red) blocks its path.

Figures 8.12 and 8.13 show the early results from the tests completed with the framework. As can be seen from Figure 8.13, the cyclist shown in blue avoided collision with the person (in red) even when its straight-line movement path was blocked.

8.2.2 Motion Influences Due to Moving Cyclist Avatars

The previously described framework explains how intelligent cyclist agents who would move along collision-free paths could be realised and visualised using mixed reality. However, for such virtual agents to enhance safety and influence pedestrian walking, their visual presence should prompt persons to maintain safety distances as in real pedestrian-cyclist interactions. Taking a traffic crossing as an example, if people interact with the virtual cyclist safely while intersecting its path, this could be an indication that persons are navigating while taking into account the presence and motion of these virtual agents. On the contrary, avoiding any interactions and walking past a cyclist avatar would mean ignoring their presence. Hence to further estimate the walking influences of a person to virtual traffic agents in the scene, this section focused on a mixed reality user study with virtual cyclists crossing walking pedestrian paths. One of the objectives of the study was to understand whether the walking path of a person would change when seeing a virtual cycle moving and crossing him.

For the MR study, the framework that was proposed in the previous section has been further extended and modified for a pilot user study. For this, rather than having cyclists as intelligent agents with collision avoidance; we program the virtual cycle to move with a constant velocity and fixed path like the crossing confederate in the experimental study detailed in 8.1. Then a participant who is expected to cross paths with the cyclists might perform a conflict avoidance maneuver to avoid risking collisions. The study in this section focuses on capturing such interaction manoeuvres between a walking person and a virtual cyclist crossing his motion path. For this, we used our mixed reality framework, the Hololens and improved the Unity application for the study that has been detailed in the subsequent sections.

8.2.3 Experimental Design and User Study

The user study consisted of a walking experiment where study participants moved in an indoor setting while wearing the Hololens and crossing a moving virtual cyclist. The crossing behaviour of the person then results in a virtual interaction with the cyclist which is termed as vInteraction in this thesis. As shown in Figure 8.14, each participant was expected to complete a journey by walking from a predefined start-point to the end-point in the lab. Based on what the participants saw while walking along the path, we designed the experiment to test three conditions:

- **no AR:** In this condition, the participant was shown no virtual mixed reality content while walking from A to B.
- **AR w\o vInteraction:** For this setting, both a virtual bench and tree were added to the left and right of the walking path of the participant. This was visible only to the participant as he walked with the MR headset on the path.
- **AR with vInteraction:** In this last condition each participant saw both static infrastructure and the 3D moving cyclist. The bench and tree along with the crossing cyclist were shown in the Hololens. The motion path for the cycle was scripted to start from a fixed start point and cross paths with the straight line path of the participant. Also, the speed for the virtual cycle was fixed and moved continuously without stopping for each of the tests.

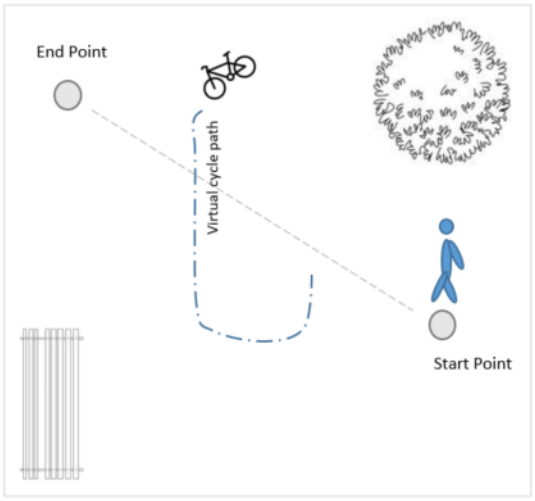


Figure 8.14: Experimental scene from a top view perspective with the cyclist crossing the walking path and the bench and the tree to the left and right respectively.

Furthermore, as the study was a walking experiment, we conducted three trails for each of the conditions with each trail based on the speed of pedestrian walking- **slow, medium** or **fast**. Hence based on the presence of AR content (condition) and speeds (trails), we designed a 3 (AR: no AR vs AR w\o vInteraction vs AR with vInteraction) X 3 (speed: slow vs normal vs fast speed) within- group study that resulted in 9 iterations in total (slow speed & no AR, slow speed & static AR, slow speed & static + dynamic AR, medium speed & no AR, medium speed & static AR, medium speed & static + dynamic AR, fast speed & no AR, fast speed & static AR, fast speed & static + dynamic AR). Our experiment for the study was conducted in an open space 10m X 10m indoor lab with optical motion tracking to track the movement of the study participants.

Implementation: As both the movement characteristics and the visual appearance of the virtual cycle were important aspects of a successful mixed-reality study, we first created a real cyclist movement dataset where a single bicycle moved in an open outdoor space and captured its data. Then using the bike motion parameters derived from the dataset, we designed a Unity cyclist GameObject for bicycle movement. Finally, the virtual cyclist asset based on the design was imported into the user study application and moved along a predefined path for the experiment.

Real Bicycle Dataset: A bicycle 3D point cloud dataset was recorded at 10 fps using a static laser scanner (Hesai PandaXT 32) observing a moving cyclist in an open space outdoor scene. Following the data capture, using a manual segmentation approach, the movement path of the bicycle was later extracted in post-processing. The rate of change of movement of the cycle along the ground plane was then used to estimate the heading angles for each time stamp in the data (Figure 8.15).

Virtual Cyclist Design: The movement information extracted from the bicycle dataset consisted of the cyclist's position and heading. Using these parameters, a 3D bicycle movement model was designed for the Unity asset. As depicted in Figure 8.16, the Unity gameObject

Figure 8.15: Figure on the left indicates the raw pointcloud (with the cyclist) captured using the 3D scanner and the right figure shows the extracted path and the rate of change of heading (indicated by tangential lines) for each time stamp.

design for the bike was a simplified motion representation (excluding the rider). In the design, given the position and heading data from the dataset, only the movement of the center of mass (CO)and the tilting angle(ϕ) of the front wheel were used to control movement in 3D. Hence at any given time point, the motion of the cycle was represented by P where P(t) =(CO, ϕ). Also for our proposed model, the tilting angle of the bicycle frame along the vertical axis (which occurs when a bicycle turns) and other more complex representations were not considered.

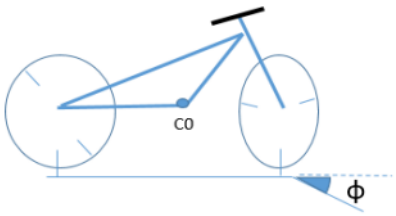


Figure 8.16: Cyclist GameObject model for animation in Unity

Following the design step, a 3D cyclist that followed the design specifications was imported to Unity and integrated into the mixed reality framework. All the other virtual 3D objects used in the study (benches and trees) were mainly freely available packages from the Unity Asset store [4]. To track the walked positions of each study participant, other than the MRTK camera positions that were transmitted by the Hololens, optical markers were placed on the Hololens and the body of each participant. This was then used to track each participant's motion externally at 200 Hz. During the study, before each participant started a trial, the mixed reality application was launched on the Hololens and scanned for QR codes placed in the indoor lab. The headset was only used in all the trails that included virtual content and interactions.

Before the experiment, the participants were briefed about the walking task to be performed. Persons were instructed to walk along the straight line path (Figure 8.14) while freely deciding to react to the virtual content if present based on the iteration conditions. They were also informed that the cyclist was not intelligent and hence would not react to the participants if they blocked its movement path. All participants were expected to complete all nine iterations including the three AR conditions with slow, medium and fast walking speed trails in random order. In the study, each participant was informed to commence

[4]https://assetstore.unity.com/

Figure 8.17: Mixed reality view of the crossing cyclist as seen from the Hololens. The figure shows the walking path along with the virtual tree that was added to the scene.

walking based on their natural slow, medium and fast speeds when an auditory beep signal was broadcasted by the experimenter. At the same instance, the cyclist started moving for all trails of the AR with vInteraction condition. A volunteer was responsible for recording all the motion trajectories of the experiment at the secondary interface.

The pilot study for this work was completed with 5 participants (2 females and 3 males) who were mostly master students or university employees. All the participant were in their mid-twenties and early thirties and had normal or corrected-to-normal vision.

8.2.4 Data Analysis

For each participant, the motion data captured in the experiment included walking trajectories recorded from two sources: a) the external motion tracker at a high frame-rate and b) the Hololens camera motion from Unity that is recorded on the secondary interface server. While the first data source provides a high-frequency movement data of how each person walked; the second source is used as a reference in correctly estimating cyclist paths with respect to the external tracking system. This is primary because both the Hololens camera and the virtual cyclist move in a fixed Unity cordinate system of the MR application. This low frequency data from virtual environment would be prone to noise and motion drifts due to MR capture. Hence by using motion clues of the person from the Hololens movement tracks that capture the same walk as the external motion tracker, the high-frequency data of the optical tracker can be spatially aligned to get a noise-free movement trajectory of all interactions recorded. Furthermore, any time lags or differences in the data can be corrected in the post-processing step. Hence following the above-mentioned alignment step we get high-frequency motion tracks for all participants for all conditions tested in the study.

The three conditions tested in the motion experiment - noAR, AR w\o vInteraction, AR with vInteraction are investigated separately in the analysis step. For the first two conditions, we only compute the speed profiles for all the participants and apply statistical analysis on the maximum walking speed. For the AR with vInteraction case, along with max speed, we also compute the three key metrics - *the Post Encroachment Time -PET*, *the euclidean distance* and *min distance* to approximate both the level of safety and the involved participant motion dynamics in the virtual interaction.

To spatially characterise the same interaction, we compute the *euclidean distance* between the two agents when the PET was estimated. This represented how far the two agents were from each other when a conflict was missed. The last metric *min distance*, represents the point of minimum separation between the person and the cyclist throughout the virtual interaction experiment. This distance value could have been at its minimum at any point after having started to walk and before the participant reached the stop point. This value was computed by replaying the path of both the cyclist and pedestrians as in the experiment and recursively estimating their distance. The computation was stopped when the point of min distance was reached for the experiment iteration.

The observations of 45 iterations (9 iteration x 5 participants) were analyzed. Furthermore, the data from one of the iteration was discarded during post-processing, mainly due to time synchronization issues.

8.2.5 Results and Discussion

In the following section, the results in terms of walking and interaction behaviour are reported. Also due to the pilot character of the study, only descriptives are reported.

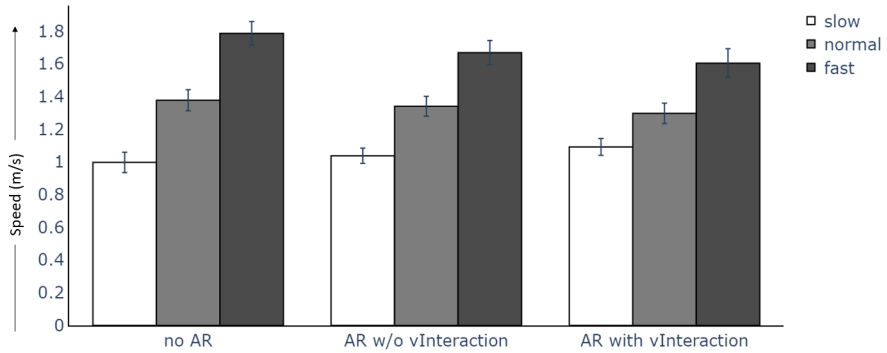


Figure 8.18: Max speed variations for no AR, AR w\o and with vInteraction for slow (white), medium (gray) and fast (black) walking speeds. The error bars represent the confidence intervals for the max speeds.

Maximum Walking Speed: As illustrated in Figure 8.18, participants were following the instructions on different walking speeds in the three speed conditions. Specifically, participants were walking slower and faster than normal in the slow and fast walking conditions, respectively. Further, the maximum walking speed does not seem to be affected by additional virtual content in the walking environments.

Interaction Behaviour: For the *AR with vInteractions* condition, analysis was applied for the crossing order to understand the choice of the interaction behaviour upon pedestrians seeing a virtual cyclist crossing their walking path. Of the total 44 trials, participants were giving way to the cyclist in 3/4th of the trials, while decided to pass first in 1/4th of the trials. As role attribution has been shown to result in different collision avoidance strategies (Olivier et al. 2013), separate descriptives are reported for all further analyses.

Table 8.3: The interaction metrics computed based on different crossing orders- cyclist first (vInteraction$_{CP}$) or pedestrian first (vInteraction$_{PC}$).

Interaction metric (mean)	vInteraction$_{CP}$	vInteraction$_{PC}$
PET (sec)	2.23	1.68
Euclidean distance (m)	2.11	1.37
Minimum distance (m)	0.995	0.809
cases (in %)	3/4	1/4

Table 8.3 illustrates the mean of the temporal and spatial crossing metrics. It can be note that values for all three metrics PET, Euclidean and Minimum distance were larger when pedestrians gave the right of way to the virtual cyclist as compared to when the cyclist crossing first. Moreover it was noted that giving way to the virtual cyclist, a strategy chosen in the majority of trials had resulted in larger temporal and spatial crossing distances between the cyclist avatar and participant.

Table 8.4: Speed and participant wise categorisation of trials with vInteraction$_{PC}$ crossing strategy

	slow	normal	fast
Total Trial Count	1	3	7
Trial Count per Participant	-	1 for P1	3 for P1
	1 for P2	2 for P2	3 for P2
	-	-	1 for P3

For being able to further explore the factors that contributed to the selection of a pedestrian crossing first-strategy, separate analyses were performed for all 11 vInteraction$_{PC}$ pedestrian first trials.

As can be noted from Table 8.4, the selection of a crossing first-strategy seems to be influenced by both walking speed and individual preferences of single pedestrians. The majority of trials (7 out of 11) can be assigned to the fast walking speed condition. Further, the crossing first-strategy was mainly selected by two pedestrians (10 out of 11 trials) only.

As the speed of walking could be an attribute that might have been specific to each of the participants P1 and P2, the maximum walking speed was computed for the *fast* iterations and compared to the speeds of participants P3-P5. It can be noted that P1 tended to walk faster than the average of P3-P5 (Figure 8.19), while the opposite seems to be true for P2. Still, P2 was able to cross in front of the virtual cyclist. This has been an interesting early finding of the work, which implies that not just speed but other factors seem to have contributed towards the crossing decisions made.

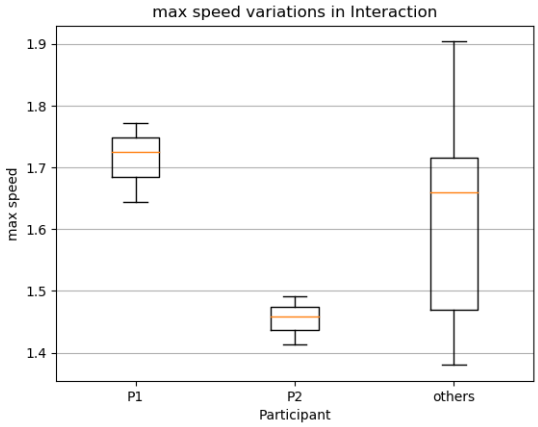


Figure 8.19: Max speed and its spread representation when comparing the speed of P1 and P2 to that of the other participants (P3 to P5) for fast-paced walking trials.

An interesting finding of the work was that both types of crossing behaviours - either crossing first or giving way could be observed as a part of the collision avoidance strategies when interacting with virtually represented cyclists. While previous research (Olivier et al. 2013, Knorr et al. 2016) has highlighted that giving the right of way is a more successful collision avoidance behaviour, the preferences of a few of the participants to cross first has been an interesting find of the current work. Also when looking at the PET and minimum distances that were computed when persons crossed first, lower values for these metrics indicated less safe crossing than when the virtual cyclist crossed. From a safety perspective, this find indicates that pedestrian crossing was safer when crossing first. We hypothesize that this could be due to the high level of predictability in estimating cyclists' motion in the present study. Once the participants estimated the potential risks, they decided to cross quickly which could have led to such unsafe behaviours. As described above, the virtual cyclist behaviour was scripted to move using a constant velocity and was not programmed to be reactive to pedestrians walking (Kamalasanan et al. 2022b) e.g. to slow down if a pedestrian approached closer.

Furthermore, the third interaction metric, i.e. *min distance*, yielded nearly similar values for both passing strategies, with *min distance* being only slightly lower when pedestrians were crossing first. This metric, unlike the Euclidean distance, was computed to estimate how close the two agents (pedestrian and virtual cyclist) were passing each other throughout the experimental trials. Taking all three metrics together, the outcomes of our study suggest that, under high predictability of the virtual cyclist's behaviour, even when being non-reactive, participants can predict the temporal and spatial distance to the cyclist and flexibly choose collision avoidance behaviour accordingly. Importantly, multiple factors, such as e.g. maximum walking speed but also pedestrian-specific preferences, seem to influence the actual choice of collision avoidance

As the primary objective, the study successfully investigated how pedestrians react to the AR content of seeing a cyclist cross their path and whether this influences their decision to either cross first or give way. While the pilot study demonstrated that walking persons preferred to interact when crossing the path of the cyclist, further investigation needs to be done to understand if virtual cyclists could be used to represent the motion dynamics of real pedestrian-cyclist interactions. If future works can accurately demonstrate the behaviours

of mixed reality cyclists to influence behaviours, then such visual approaches could be used to improve walking and crossing visibility in partially occluded situations.

9 Discussion and Outlook

While the results for each of the sections have been evaluated individually, in this chapter, the findings from the different methods are compared and further discussed for potential future development. The chapter starts with an overview, followed by an assessment of the motion perception capability of the AR device and discusses its performance evaluation. Then the effects of showing the future path in AR and how this might influence path choices is discussed. The chapter concludes with suggestions for future research identified based on the results, that needs to be addressed in the future.

This thesis investigated the effects of using an AR device and its visual interface to influence the walking behaviour of pedestrians. The methodology involved demonstrating a motion pipeline to perceive motion in the ego view of the Hololens and studying the effects of showing the future motion paths. Pedestrian safety issues while walking in shared spaces was focused to be addressed using the proposed method of AR influence. For this a motion perception algorithm used the data from the Hololens and first detects and then tracks walking pedestrians in the field of view. For this, the thesis compares applying particle filter tracking to both F-Pointnet and MaskRCNN based pedestrian detection. Also the preferred path choice differences of seeing the future motion of others in front of the Hololens is studied. The future motion path is represented in AR either with a "simple 3D arrow" or with an "confident" or "uncertain" encoded arrow. The different designs used for encoding are based on visual variables commonly used in cartography and map representations.

To evaluate on how well the motion pipeline detected and tracked pedestrians, both F-Pointnet and MaskRCNN based 3D detections and tracked paths were measured using the 3D IoU and CLEAR Mot metric respectively. Furthermore the walking influences of seeing future motion of others were studied from a safety perspective; using walking path preferences, while crossing paths with persons whose future trajectory was visualised with AR. The surrogate safety measure PET was used to estimate whether persons preferred to walk longer or shorter paths when seeing the future information to cross.

Finally other influences that could result in walking when using AR were also experimented in the work. When walking behaviours were controlled using a virtual traffic infrastructure, people reacted to crossing conflicts with more safety and higher TTC. Also, it was noted that walking behaviours were more predictable when AR mediated crossing priority. For the methods that tested crossing behaviours with cyclists as virtual avatars, different behaviours based on who crossed first was noted. When persons crossed the cyclist first, less safe interaction with lower PET values were noted than for all those situations when cyclists crossed first.

9.1 Motion Perception using AR Hololens

Overall, the results presented in this thesis have proven that it is possible to both detect and track pedestrians using the RGBD sensors of an AR device. This can then be used to apply visualisations based on interpretations of surrounding motion with AR.

In general, even when the findings from the work verify motion perception using two detection approaches - Frustum Pointnet and MaskRCNN for the Hololens, both the quantitative and qualitative results need to be investigated further. This is because, for AR influence, both detection and tracking are important. The detection outputs - the centroid of the 3D box that localises the position and the 3D shape (via bounding dimensions) of a person. The tracking step estimates - the walked paths of the person by inputting 3D positions from the detector and associating them for the different time stamps. The errors made in correctly detecting and tracking persons can influence effective visualisation using its results.

The 3D object detection accuracy (AP) metric when testing pedestrian detections with F-Pointnet (without 2D pose features) for the simulated shared space dataset (SSS) yielded a score of 0.89 (IoU=0.3). This score was 32.8 % higher than the 3D detection accuracy for MaskRCNN detection using the same dataset. This indicated that the pedestrian detection using the AR device Hololens was able to detect the approximate body shapes of persons and their orientations better in 3D using the F-Pointnet. However, when the real-time implementation of F-Pointnet and MaskRCNN were evaluated for tracking performance with the IKG pedestrian tracking dataset, the MaskRCNN tracks performed better than F-Pointnet for the CLEAR MOT metric. The MaskRCNN in our work showed a better tracking accuracy of 81.39 % which was higher than F-Pointnet by 10%. The better performance of F-Pointnet in 3D detection could indicate towards the network overfitting to the SSS Dataset while a higher tracking performance of MaskRCNN could mean better generalisation of the tracking-by-detection MaskRCNN pipeline. However both these claims need to be verified with further investigation.

The MOTP score computed for MaskRCNN clearly indicated that the average pedestrian localisation error was under 20 centimeter. This was reported to be considerably lower than the F-Pointnet-result, where errors were as large as 40 cm. Hence our MaskRCNN pipeline reported a relatively good localisation performance while person tracking. If the future path of others have to be visualised, the pedestrian localisation accuracy approximated is acceptable considering the definition of position uncertainty as a person walks. This has been further elaborated in the next paragraph. Also, the CLEAR MOT scores obtained for the IKG Pedestrian Tracking Dataset largely depends on how pedestrians moved in the scene and how well the tracker performed (Nguyen 2020). Future work needs to verify whether the tracker performance recorded in this thesis are comparable to well performing 3D pedestrian trackers like OC-SORT (Cao et al. 2022) that are evaluated with the KITTI benchmark (Geiger et al. 2015). Also, our work has only focused on using the results from the detector directly to track pedestrians with it. All the detected outputs of the pedestrian detector, including false positives, are taken as input to the tracker. Future work could consider adding a preprocessing step as proposed in (Nguyen 2020), where noisy inputs are filtered out before tracking pedestrians with it.

Furthermore, if the output of the motion perception pipeline is to be used for AR visualisation to support safety, then the computed MOTA and MOTP metrics have to be evaluated from both AR and safety perspective, respectively. This would then require that the motion perception pipeline has a good (a) localisation and (b) tracking with acceptable id-switching. The localisation accuracy in this case is important because any visualisation that is placed at the position of each tracked person should be accurate to ensure the registration of both real and accurate virtual content to influence motion. Misalignments between the virtual and real content could result in perception issues and ineffective visual influences (Drascic and Milgram 1996). In this regard, errors from the trackers as high as 20 cm are still accept-

able to show either simple or informative future. The uncertain future path when visualised with the pipeline then will have an error bound within a range of ± 20cm.

On the contrary, the consistency of tracked id returned for a person in the scene is only important if the visualisation used gives a color scheme or visual attribute to differentiate each person. However our future path visualisation only prioritizes to show the future positions and whether the predictions are confident or uncertain. This further emphasises on revising the use of the MOTA metric as accuracy score of the tracker. The missed tracks and false positives still need to be considered without prioritising id-switches that happened while tracking. Future work could consider the use of the HOTA metric (Luiten et al. 2021), that is a successor of the CLEAR MOT metric which might fit better to tracking future path based AR motion.

Lastly when different methods that are proposed to influence walking have been measured using the surrogate safety measures (PET and TTC) in this thesis, the motion perception pipeline has not been evaluated from a safety perspective. Some recent works have demonstrated early results in evaluating computer vision algorithms from a safety perspective. Lyssenko et al. (2021) evaluated pedestrian detection using the IoU to see how they performed in situations when sensors moved in close proximity to walking persons. Using a simulated dataset from autonomous driving, pedestrians walking closer to the sensor and their localisation performance were prioritised over others in the scene. The performance of the evaluated detector was then proven to reduce in precision with increasing distance. To evaluate perception using 3D sensors differently, Wolf et al. (2021) proposed a safety aware people detection metric that took into account both the distance of the person from the sensor and also the resulting time-to-collision (TTC). This metric metric was then used to provide meaningful assessments in safety situations. The use of a surrogate safety metrics like TTC have further been evaluated for other perception performance works (Lyssenko et al. 2022, Philion et al. 2020). While TTC and PET have been used to study AR influences from a safety perspective in our work, future works could also consider incorporating it directly into the motion perception evaluation step. This might get a better idea of how our pipeline estimated PET and the severity of the resulting motion conflicts.

9.2 Walking Influence Based on Scene Motion Visualisation

Future motion information is considered to be advantages in safety critical scenarios where visual information would be provided to either warn or assist pedestrians in making safer path choices. Some recent works have informed pedestrians or vehicle occupants about either occlusions and resulting dangers or about future motion (Peereboom et al. (2023), Colley et al. (2017a)). The motion influences that result from such works are more focused for specific use cases (like moving behind parked cars or improving trust in machines). Our future path visualisation approach presented in this thesis could be extended to influence pedestrians in different safety critical situations that could include crossing paths, or paths that might cross in a future point in time.

In our work, the path choice effects of showing the future path were studied both for a simple and an informative (confident or uncertain) arrow based 3D visualisation. In each of the cases, the tip of the arrow indicated the position of the person in the future and its orientation indicated how s/he rotated the body or changed the direction of walking. All AR prediction visualisations were shown as a projection on the ground. For the majority of

test participants who decided to cross from the front proved safer walking to result from the use of AR visualisations. This held for all scenes irrespective of their inherent complexity to guess the future. However when the prediction information of either "confidence" or "uncertainty" were encoded in AR, then the participants showed mixed behaviour. Scenes that were less difficult to predict future motion prompted less safety conscious crossing behaviours and vice versa for high complexity scenes. All safe walking routes were noted to result in longer crossing paths in the study.

The methods to evaluate safety behaviours that resulted from future path visualization was extensively studied based on how persons choose to cross from the front or from behind. This helped to understand both simple future and informative future paths in AR. The study pointed out that the path choice behaviours to safety cross differed based on whether a person crossed from the front or from behind. However, this crossing based analysis was proven to be limited by the size of the collected participant data as for few scenes only less participants crossed from the behind. Future work could consider the imbalance of crossing samples when designing the experiments.

To summarise the findings from this thesis, the work successfully demonstrates an ego pedestrian perception algorithm that detects and tracks other walking pedestrians in front of the Hololens. Furthermore, the work also proves that AR can influence walking based on future path visualization. Along with this contribution, the Chapter 8 presented two different methods of walking influences using virtual content to represent static infrastructure or dynamic scene motion. The safety improvements when walking with adhoc virtual signals that guide people to stop or cross during a crossing situations was studied in the first. In the second method, moving cyclists as virtual 3D traffic avatars were tested in a walking experiment with pedestrians. The findings of the study explored how crossing behaviours and interactions differed when persons crossed virtual moving cyclists in mixed reality. While the findings of the first study can be useful to traffic planner and urban designer to install virtual traffic systems, the learning's from the second method can be used to understand the benefits of represent moving agents in used cases like occlusion due to parked cars or limited Line-Of-Sight (LOS) visibility.

9.3 Suggestions for Future Work

To influence the walking behaviour using AR, this thesis has restricted itself to the Hololens and proved AR influences to pedestrian walks in indoor or space constrained settings. This in turn indicates that there is a potential for further research questions related to the overall research topic. Therefore this chapter provides a selection of suggestions for future research directions that are not addressed in the scope of this thesis.

While this thesis provides results focused on shared spaces and influencing path choices during collisions with both motion perception and AR, the findings from the work could also be used to address safety risks in other domains. This, for example, could include *industrial co-working spaces* where humans would share spaces with robots or *future connected urban environments* where people share space with interconnected autonomous vehicles. In the former use case, visualisations shown with AR headsets would guide pedestrians to walk when being confronted with a mixed group of robots and humans in a factory setting. If AR were to be used in a future urban scenario as mentioned in the latter case, headsets available in smaller form factors and higher technical capability could further traffic com-

munications. This might then make it possible for pedestrians to directly talk to a groups of autonomous vehicles. If vehicles are informed about pedestrian crossing urgency, newer forms of negotiations could result in traffic spaces (Li et al. 2022b).

The perception methodology proposed in this thesis is strongly based on the assumption that the Hololens is positioned at a stationary point in the scene and its sensors are recording the RGB and depth of the surrounding movement. The designed visual interface to show the future path of others also has not studied non stationary ego user scenes. While with a static AR headset it was possible to prove the concept of AR influence, challenges that would arise when considering the device to move with walking needs to be addressed. Two research gaps have to be primarily addressed within the scope of AR influence for this. Firstly, as the headset would be in motion, all tracked positions of other walking pedestrians have to be now estimated in a global coordinate system. This would require localising the headset in a global reference frame, estimating the pose and then transforming detected tracks to this global frame. Secondly, when visualisations are applied to a moving headset, there could be dynamic errors that have to be accounted for. The first error in this regard is the localisation error which might result in wrongly estimated global pose which might cause virtually placed holograms to drift or float (Vassallo et al. 2017). The second issue that would need attention is the amount of motion estimation uncertainty that would arise when applying scene motion perception with relative movement. The current finding of the thesis have reported low errors in tracking walked persons when using a MaskRCNN and a particle filter for the Hololens. This would have to be reevaluated as headset movement could introduce additional pose errors when tracking surrounding pedestrians. This might further affect visual experiences when showing the future path of other walking nearby with AR.

As a part of the proposed method, the motion perception component and future path visualisation component have been considered separately for the application of the methodology. However, there is a possibility to apply the proposed methodology as a combination of both. This would then require the scene motion pipeline to also estimate future trajectories along with both detecting and tracking pedestrian movement. Hence research on estimating future path in real-time for the Hololens has to be explored. This would require the implementation of tracking in real-time to design the amount of frames that would be required to predict immediate future. Also while there has been significant research within the robotic community on motion prediction, methods that have applied pedestrian motion prediction for shared spaces (Cheng et al. 2023, Hossain et al. 2022) might be worth exploring further.

Also motion predictions could benefit from more sensor information about the person. For example, as a predicted path would depend on the goal/destination, level of motivation, etc. This information could be better understand or interpret with more sensor data. As the person would wear an AR Hololens, it might be possible to use its sensors to create internal models that could help to explain psychological traits about the AR user. For example by using data from sensors (eye tracking for instance) to understand gaze, it might be possible to infer the level of attention, estimate fatigue and also approximate his/her intended destination. This can then be used to create a digital twin of the AR headset user such that motion influences from AR could be better interpreted with the twin model.

As the thesis does not cover conducting a usability test to investigate the effectiveness of the method, future works could focus on replicating the finding of this thesis. For this, the complete pipeline and the future visualisation explained in this thesis have to be implemented in real-time. Future works in this direction should investigate the usability of a

fully functional online system and also address the differences in the effectiveness of using AR by comparison to the user studies that are conducted as a part of this thesis.

Finally, even when virtual cyclists have been proposed as a visualisation method to influence walking, the use case in improving vision of occluded cyclists with MR needs to be tested. Future research could investigate to visualise occluded cyclists to support safety of vulnerable road users in upcoming works.

Scenario Specific Visualisations of Future Motion Path: In this thesis, the methodology to influence walking with future path visualization (Section 7) and other influences (Section 8) has been tested for single pedestrian scenarios. However, as walking of neighboring persons always tends to directly influence others in walking, future work should focus on how visual influences would be different considering larger scenes and more interactions between different pedestrians. It might also be worth exploring whether the use of the arrow visualisation to indicate future paths would create visual clutter in the scene.

In this context, the use of different graphical variables other than arrows, e.g. colours and symbols could be investigated in subsequent works. It might be interesting to observe whether the use of animations, e.g. a glowing circle instead of a moving arrow, would give the same level of crossing influences.

List of Figures

List of Tables

Bibliography

Ackermann, Claudia, Matthias Beggiato, Sarah Schubert, and Josef F Krems (2019). "An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles?" In: *Applied ergonomics* 75, pp. 272–282.

Ahmed, Suhair, Fatema T Johora, and Jörg P Müller (2020). "Investigating the role of pedestrian groups in shared spaces through simulation modeling". In: *Simulation Science: Second International Workshop, SimScience 2019, Clausthal-Zellerfeld, May 8-10, 2019, Revised Selected Papers 2*. Springer, pp. 52–69.

Albarrak, Luluah, Oussama Metatla, and Anne Roudaut (2019). "An Exploratory Study for Evaluating the Use of Floor Visualisations in Navigation Decisions". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–6.

– (2020). "Exploring the Design of History-Enriched Floor Interfaces for Asynchronous Navigation Support". In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pp. 1391–1403.

– (2021). "(don't) mind the step: Investigating the effect of digital social cues on navigation decisions". In: *Proceedings of the ACM on Human-Computer Interaction* 5.ISS, pp. 1–18.

Alfakhori, Muhammad, Habiburrahman Dastageeri, Sven Schneider, and Volker Coors (2022). "Occlusion screening using 3d city models as a reference database for mobile ar-applications". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 10, pp. 11–18.

Alink, GMM (1990). "Road Safety Policy In The Netherlands and the Effect of the Infrastructure on the success of the Policy". In: *Living And Moving In Cities. Proceeding Of The Congress, Paris, January 29-31*.

Allen, Brian L, B Tom Shin, and Peter J Cooper (1978). *Analysis of traffic conflicts and collisions*. Tech. rep.

Andersson, Dan, Lina Wahlgren, and Peter Schantz (2023). "Pedestrians' perceptions of route environments in relation to deterring or facilitating walking". In: *Frontiers in Public Health* 10, p. 1012222.

Appleyard, Donald (1980). "Livable streets: protected neighborhoods?" In: *The ANNALS of the American Academy of Political and Social Science* 451.1, pp. 106–117.

Aschermann, Malte, Philipp Kraus, and Jörg P Müller (2016). "LightJason: a BDI framework inspired by Jason". In: *European Conference on Multi-Agent Systems*. Springer, pp. 58–66.

Avila Soto, Mauro and Markus Funk (2018). "Look, a guidance drone! assessing the social acceptability of companion drones for blind travelers in public spaces". In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 417–419.

Bahri, Haythem, David Krčmařík, and Jan Kočí (2019). "Accurate object detection system on hololens using yolo algorithm". In: *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*. IEEE, pp. 219–224.

Bark, Karlin, Emily Hyman, Frank Tan, Elizabeth Cha, Steven A Jax, Laurel J Buxbaum, and Katherine J Kuchenbecker (2014). "Effects of vibrotactile feedback on human learning of arm motions". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23.1, pp. 51–63.

Batista, Mariana and Bernhard Friedrich (2022a). "Analysing the influence of a farmers' market on spatial behaviour in shared spaces". In: *Journal of Urban Design* 27.5, pp. 528–545.

– (2022b). "Investigating spatial behaviour in different types of shared space". In: *Transportation research procedia* 60, pp. 44–51.

Bazilinskyy, Pavlo, Dimitra Dodou, and Joost De Winter (2019). "Survey on eHMI concepts: The effect of text, color, and perspective". In: *Transportation research part F: traffic psychology and behaviour* 67, pp. 175–194.

Berclaz, Jerome, Francois Fleuret, Engin Turetken, and Pascal Fua (2011). "Multiple object tracking using k-shortest paths optimization". In: *IEEE transactions on pattern analysis and machine intelligence* 33.9, pp. 1806–1819.

Bernardin, Keni and Rainer Stiefelhagen (2008). "Evaluating multiple object tracking performance: the clear mot metrics". In: *EURASIP Journal on Image and Video Processing* 2008, pp. 1–10.

Berton, Florian, Anne-Hélène Olivier, Julien Bruneau, Ludovic Hoyet, and Julien Pettré (2019). "Studying gaze behaviour during collision avoidance with a virtual walker: Influence of the virtual reality setup". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, pp. 717–725.

Bimber, Oliver, Daisuke Iwai, Gordon Wetzstein, and Anselm Grundhöfer (2008). "The visual computing of projector-camera systems". In: *ACM SIGGRAPH 2008 classes*, pp. 1–25.

Bimber, Oliver and Ramesh Raskar (2005). *Spatial augmented reality: merging real and virtual worlds*. CRC press.

Blue, Victor J and Jeffrey L Adler (1999). "Cellular automata microsimulation of bidirectional pedestrian flows". In: *Transportation Research Record* 1678.1, pp. 135–141.

Boarnet, Marlon G, Kristen Day, Craig Anderson, Tracy McMillan, and Mariela Alfonzo (2005). "California's Safe Routes to School program: impacts on walking, bicycling, and pedestrian safety". In: *Journal of the American Planning Association* 71.3, pp. 301–317.

Bolte, Benjamin and Markus Lappe (2015). "Subliminal reorientation and repositioning in immersive virtual environments using saccadic suppression". In: *IEEE transactions on visualization and computer graphics* 21.4, pp. 545–552.

Bouguet, Jean-Yves (2004). "Camera calibration toolbox for matlab". In: *http://www. vision. caltech. edu/bouguetj/calib_ doc/.*

Bruder, Gerd, Phil Wieland, Benjamin Bolte, Markus Lappe, and Frank Steinicke (2013). "Going with the flow: Modifying self-motion perception with computer-mediated optic flow". In: *2013 IEEE International symposium on mixed and augmented reality (ISMAR)*. IEEE, pp. 67–74.

Buck, Lauren E, John J Rieser, Gayathri Narasimham, and Bobby Bodenheimer (2019). "Interpersonal affordances and social dynamics in collaborative immersive virtual environments: Passing together through apertures". In: *IEEE transactions on visualization and computer graphics* 25.5, pp. 2123–2133.

Busch, Steffen, Alexander Schlichting, and Claus Brenner (2018). "Generation and communication of dynamic maps using light projection". In: *Proceedings of the ICA*. Vol. 1. Copernicus GmbH, pp. 1–8.

Cao, Jinkun, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani (2022). *Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking*. arXiv: 2203.14360 [cs.CV]. URL: https://arxiv.org/abs/2203.14360.

Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2017). "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Cappé, Olivier, Simon J Godsill, and Eric Moulines (2007). "An overview of existing methods and recent advances in sequential Monte Carlo". In: *Proceedings of the IEEE* 95.5, pp. 899–924.

Chen, Long, Wen Tang, Nigel John, Tao Ruan Wan, and Jian Jun Zhang (2018). "Context-aware mixed reality: A framework for ubiquitous interaction". In: *arXiv preprint arXiv:1803.05541*.

Chen, Peng, Weiliang Zeng, Guizhen Yu, Yunpeng Wang, et al. (2017). "Surrogate safety analysis of pedestrian-vehicle conflict at intersections using unmanned aerial vehicle videos". In: *Journal of advanced transportation* 2017.

Chen, Zhutian, Qisen Yang, Jiarui Shan, Tica Lin, Johanna Beyer, Haijun Xia, and Hanspeter Pfister (2023). "iBall: Augmenting Basketball Videos with Gaze-moderated Embedded Visualizations". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18.

Chen, Zhutian, Shuainan Ye, Xiangtong Chu, Haijun Xia, Hui Zhang, Huamin Qu, and Yingcai Wu (2021). "Augmenting sports videos with viscommentator". In: *IEEE Transactions on Visualization and Computer Graphics* 28.1, pp. 824–834.

Cheng, Hao, Mengmeng Liu, Lin Chen, Hellward Broszio, Monika Sester, and Michael Ying Yang (2023). "Gatraj: A graph-and attention-based multi-agent trajectory prediction model". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 205, pp. 163–175.

Cheng, Lung-Pan, Eyal Ofek, Christian Holz, and Andrew D Wilson (2019). "Vroamer: generating on-the-fly VR experiences while walking inside large, unknown real-world building environments". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, pp. 359–366.

Choi, Wongun (2015). "Near-online multi-target tracking with aggregated local flow descriptor". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3029–3037.

Chowdhury, Md Fahim, Md Ryad Ahmed Biplob, and Jia Uddin (2018). "Real time traffic density measurement using computer vision and dynamic traffic control". In: *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, pp. 353–356.

Cinelli, Michael Eric and Aftab E Patla (2008). "Task-specific modulations of locomotor action parameters based on on-line visual information during collision avoidance with moving objects". In: *Human movement science* 27.3, pp. 513–531.

Coleman, Simon and Peter Collins (2020). *Locating the field: space, place and context in anthropology*. Routledge.

Colley, Ashley, Jonna Häkkilä, Bastian Pfleging, and Florian Alt (2017a). "A design space for external displays on cars". In: *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, pp. 146–151.

Colley, Ashley, Lasse Virtanen, Pascal Knierim, and Jonna Häkkilä (2017b). "Investigating drone motion as pedestrian guidance". In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*, pp. 143–150.

Colley, Mark, Elvedin Bajrovic, and Enrico Rukzio (2022). "Effects of pedestrian behavior, time pressure, and repeated exposure on crossing decisions in front of automated vehicles equipped with external communication". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

Collins, Jonny, Holger Regenbrecht, and Tobias Langlotz (2017). "Visual coherence in mixed reality: A systematic enquiry". In: *Presence* 26.1, pp. 16–41.

Croft, James L and Derek Panchuk (2018). "Watch where you're going? Interferer velocity and visual behavior predicts avoidance strategy during pedestrian encounters". In: *Journal of motor behavior* 50.4, pp. 353–363.

Cunningham, Douglas W and Christian Wallraven (2013). "Understanding and Designing Perceptual Experiments." In: *Eurographics (Tutorials)*.

Dehghan, Afshin, Shayan Modiri Assari, and Mubarak Shah (2015). "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4091–4099.

Delaney, Hannah (2016). "Walking and cycling interactions on shared-use paths". PhD thesis. University of the West of England.

Derpanis, Konstantinos G (2010). "Overview of the RANSAC Algorithm". In: *Image Rochester NY* 4.1, pp. 2–3.

Dey, Arindam, Mark Billinghurst, Robert W Lindeman, and J Edward Swan (2018). "A systematic review of 10 years of augmented reality usability studies: 2005 to 2014". In: *Frontiers in Robotics and AI* 5, p. 37.

Dey, Debargha, Azra Habibovic, Andreas Löcken, Philipp Wintersberger, Bastian Pfleging, Andreas Riener, Marieke Martens, and Jacques Terken (2020). "Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces". In: *Transportation Research Interdisciplinary Perspectives* 7, p. 100174.

Dey, Debargha, Andrii Matviienko, Melanie Berger, Bastian Pfleging, Marieke Martens, and Jacques Terken (2021). "Communicating the intention of an automated vehicle to pedestrians: The contributions of eHMI and vehicle behavior". In: *it-Information Technology* 63.2, pp. 123–141.

Dimitrievski, Martin, Peter Veelaert, and Wilfried Philips (2019). "Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle". In: *Sensors* 19.2, p. 391.

Distefano, N, G Pulvirenti, and S Leonardi (2021). "Neighbourhood walkability: Elderly's priorities". In: *Research in transportation business & management* 40, p. 100547.

Djuric, Petar M, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez (2003). "Particle filtering". In: *IEEE signal processing magazine* 20.5, pp. 19–38.

Do, Ann (2002). "Walking the safety Walk". In: *Public roads* 66.2, pp. 2–5.

Drascic, David and Paul Milgram (1996). "Perceptual issues in augmented reality". In: *Stereoscopic displays and virtual reality systems III*. Vol. 2653. Spie, pp. 123–134.

Dubrofsky, Elan (2009). "Homography estimation". In: *Diplomová práce. Vancouver: Univerzita Britské Kolumbie* 5.

Dünser, Andreas, Mark Billinghurst, James Wen, Ville Lehtinen, and Antti Nurminen (2012). "Exploring the use of handheld AR for outdoor navigation". In: *Computers & Graphics* 36.8, pp. 1084–1095.

Evans, RW, AP Ramsbottom, and DW Sheel (1989). "Head-up displays in motor cars". In: *1989 Second International Conference on Holographic Systems, Components and Applications*. IET, pp. 56–62.

Everingham, Mark, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman (June 2010). "The Pascal Visual Object Classes (VOC) Challenge". In: *Int. J. Comput. Vision* 88.2, 303–338. ISSN: 0920-5691. DOI: 10.1007/s11263-009-0275-4. URL: https://doi.org/10.1007/s11263-009-0275-4.

Fajen, Brett R and William H Warren (2003). "Behavioral dynamics of steering, obstable avoidance, and route selection." In: *Journal of Experimental Psychology: Human Perception and Performance* 29.2, p. 343.

Feiner, Steven, Blair MacIntyre, Tobias Höllerer, and Anthony Webster (1997). "A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment". In: *Personal Technologies* 1, pp. 208–217.

Ferstl, David, Christian Reinbacher, Gernot Riegler, Matthias Rüther, and Horst Bischof (2015). "Learning Depth Calibration of Time-of-Flight Cameras." In: *BMVC*, pp. 102–1.

Fink, Philip W, Patrick S Foo, and William H Warren (2007). "Obstacle avoidance during walking in real and virtual environments". In: *ACM Transactions on Applied Perception (TAP)* 4.1, 2–es.

Fisher, MH (1930). "Optokinetisch ausgeloste Bewegungswahrnehmung und optokinetischer Nystagmus". In: *Journal fur Psychologie und Neurologie* 41, pp. 273–308.

Fu, Ting, Weichao Hu, Luis Miranda-Moreno, and Nicolas Saunier (2019). "Investigating secondary pedestrian-vehicle interactions at non-signalized intersections using vision-based trajectory data". In: *Transportation research part C: emerging technologies* 105, pp. 222–240.

Fuest, Stefan, Monika Sester, and Amy L Griffin (2023). "Nudging travellers to societally favourable routes: The impact of visual communication and emotional responses on decision making". In: *Transportation research interdisciplinary perspectives* 19, p. 100829.

Furukawa, Masahiro, Hiromi Yoshikawa, Taku Hachisu, Shogo Fukushima, and Hiroyuki Kajimoto (2011). "" Vection field" for pedestrian traffic control". In: *Proceedings of the 2nd Augmented Human International Conference*, pp. 1–8.

Gallagher, Nancy Ambrose, Kimberlee A Gretebeck, Jennifer C Robinson, Elisa R Torres, Susan L Murphy, and Kristy K Martyn (2010). "Neighborhood factors relevant for walking in older, urban, African American adults". In: *Journal of aging and physical activity* 18.1, pp. 99–115.

Geiger, Andreas, Philip Lenz, Christoph Stiller, and Raquel Urtasun (2015). "The kitti vision benchmark suite". In: *URL http://www. cvlibs. net/datasets/kitti* 2.5.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun (2012). "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 3354–3361.

Grant, Theresa L, Nancy Edwards, Heidi Sveistrup, Caroline Andrew, and Mary Egan (2010). "Neighborhood walkability: older people's perspectives from four neighborhoods in Ottawa, Canada". In: *Journal of aging and physical activity* 18.3, pp. 293–312.

Grasset, Raphael, Alessandro Mulloni, Mark Billinghurst, and Dieter Schmalstieg (2011). "Navigation techniques in augmented and mixed reality: Crossing the virtuality continuum". In: *Handbook of Augmented Reality*, pp. 379–407.

Grechkin, Timofey, Mahdi Azmandian, Mark Bolas, and Evan Suma (2015). "Towards context-sensitive reorientation for real walking in virtual reality". In: *2015 IEEE Virtual Reality (VR)*. IEEE, pp. 185–186.

Grechkin, Timofey, Jerald Thomas, Mahdi Azmandian, Mark Bolas, and Evan Suma (2016). "Revisiting detection thresholds for redirected walking: Combining translation and curvature gains". In: *Proceedings of the ACM symposium on applied perception*, pp. 113–120.

Greedy, Locally (n.d.). "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects". In: ().

Hamilton-Baillie, Ben (2004). "Urban design: Why don't we do it in the road? Modifying traffic behavior through legible urban design". In: *Journal of Urban Technology* 11.1, pp. 43–62.

– (2008). "Shared space: Reconciling people, places and traffic". In: *Built environment* 34.2, pp. 161–181.

Hannerz, Ulf (1980). *Exploring the city. Inquiries toward an urban anthropology*. Columbia University Press.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

Helbing, Dirk and Peter Molnar (1995). "Social force model for pedestrian dynamics". In: *Physical review E* 51.5, p. 4282.

Hesenius, Marc, Ingo Börsting, Ole Meyer, and Volker Gruhn (2018). "Don't panic! Guiding pedestrians in autonomous traffic with augmented reality". In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pp. 261–268.

Hirt, Christian, Marco Ketzel, Philip Graf, Christian Holz, and Andreas Kunz (2022). "Short-term Path Prediction for Spontaneous Human Locomotion in Arbitrary Virtual Spaces". In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, pp. 554–559.

Holländer, Kai, Andy Krüger, and Andreas Butz (2020). "Save the Smombies: App-assisted street crossing". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–11.

Höllerer, Tobias and Steve Feiner (2004). "Mobile augmented reality". In: *Telegeoinformatics: Location-based computing and services* 21, pp. 221–260.

Höllerer, Tobias, Steven Feiner, Tachio Terauchi, Gus Rashid, and Drexel Hallaway (1999). "Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system". In: *Computers & Graphics* 23.6, pp. 779–785.

Hossain, Sakif, Fatema T Johora, Jörg P Müller, Sven Hartmann, and Andreas Reinhardt (2022). "SFMGNet: A physics-based neural network to predict pedestrian trajectories". In: *arXiv preprint arXiv:2202.02791*.

Huang, Chuanli, Lu Wang, Hang Yu, Hongliu Li, Jun Zhang, Weiguo Song, Siuming Lo, and Warda Rafaqat (2021). "Investigating the influence of a cyclist on crowd behaviors on a shared road". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.8, p. 083402.

Huber, Markus, Yi-Huang Su, Melanie Krüger, Katrin Faschian, Stefan Glasauer, and Joachim Hermsdörfer (2014). "Adjustments of speed and path when avoiding collisions with another pedestrian". In: *PloS one* 9.2, e89589.

Hübner, Patrick, Kate Clintworth, Qingyi Liu, Martin Weinmann, and Sven Wursthorn (2020). "Evaluation of HoloLens tracking and depth sensing for indoor mapping applications". In: *Sensors* 20.4, p. 1021.

Hübner, Patrick, Martin Weinmann, and Sven Wursthorn (2018). "Marker-based localization of the microsoft hololens in building models". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 195–202.

Hydén, Christer (1987). "The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique". In: *Bulletin Lund Institute of Technology, Department* 70.

Ishii, Akira, Ippei Suzuki, Shinji Sakamoto, Keita Kanai, Kazuki Takazawa, Hiraku Doi, and Yoichi Ochiai (2016). "Optical marionette: Graphical manipulation of human's walking direction". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 705–716.

Itoh, Yuta, Jason Orlosky, Kiyoshi Kiyokawa, and Gudrun Klinker (2016). "Laplacian vision: Augmenting motion prediction via optical see-through head-mounted displays". In: *Proceedings of the 7th Augmented Human International Conference 2016*, pp. 1–8.

Jacques, Bertin (1983). "Semiology of graphics: diagrams, networks, maps". In: *University of Wisconsin Press, Madison, Wisconsin.*

Jansen, Sander EM, Alexander Toet, and Peter J Werkhoven (2011). "Human locomotion through a multiple obstacle environment: strategy changes as a result of visual field limitation". In: *Experimental Brain Research* 212, pp. 449–456.

Jensen, Ole B (2010). "Negotiation in motion: Unpacking a geography of mobility". In: *Space and culture* 13.4, pp. 389–402.

Johnsson, Carl, Aliaksei Laureshyn, and Tim De Ceunynck (2018). "In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators". In: *Transport reviews* 38.6, pp. 765–785.

Johora, Fatema T and Jörg P Müller (2020). "Zone-specific interaction modeling of pedestrians and cars in shared spaces". In: *Transportation research procedia* 47, pp. 251–258.

Johora, Fatema Tuj (2022). *Modeling Interactions Among Pedestrians and Cars in Shared Spaces.* Springer.

Jovancevic-Misic, Jelena and Mary Hayhoe (2009). "Adaptive gaze control in natural environments". In: *Journal of Neuroscience* 29.19, pp. 6234–6238.

Kamalasanan, Vinu, Yu Feng, and Monika Sester (2022a). "Improving 3d pedestrian detection for wearable sensor data with 2d human pose". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; V-4-2022* 4, pp. 219–226.

Kamalasanan, Vinu, Melanie Krüger, and Monika Sester (2023). "Developing a Cyclist 3D GameObject for a Mixed Reality Interaction Framework". In: *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).* IEEE, pp. 254–256.

Kamalasanan, Vinu, Awad Mukbil, Monika Sester, and Jörg P Müller (2022b). "Mixed reality agent-based framework for pedestrian-cyclist interaction". In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct).* IEEE, pp. 363–368.

Kaparias, Ioannis, Michael GH Bell, T Biagioli, L Bellezza, and B Mount (2015). "Behavioural analysis of interactions between pedestrians and vehicles in street designs with elements of shared space". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 30, pp. 115–127.

Kari, Mohamed, Tobias Grosse-Puppendahl, Luis Falconeri Coelho, Andreas Rene Fender, David Bethge, Reinhard Schütte, and Christian Holz (2021). "Transformr: Pose-aware object substitution for composing alternate mixed realities". In: *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* IEEE, pp. 69–79.

Kim, Dahun, Sanghyun Woo, Joon-Young Lee, and In So Kweon (2019). "Deep video inpainting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5792–5801.

Kim, Hyungil, Jessica D Isleib, and Joseph L Gabbard (2016). "Virtual shadow: making cross traffic dynamics visible through augmented reality head up display". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* Vol. 60. 1. SAGE Publications Sage CA: Los Angeles, CA, pp. 2093–2097.

Kiyokawa, Kiyoshi (2016). "Occlusion displays". In: *Handbook of Visual Display Technology*, pp. 1–9.

Kjellin, Andreas, Lars Winkler Pettersson, Stefan Seipel, and Mats Lind (2008). "Evaluating 2D and 3D visualizations of spatiotemporal information". In: *ACM Transactions on Applied Perception (TAP)* 7.3, pp. 1–23.

Klinger, Tobias, Frank Rottensteiner, and Christian Heipke (2017). "Probabilistic multi-person localisation and tracking in image sequences". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 127, pp. 73–88.

Knierim, Pascal, Steffen Maurer, Katrin Wolf, and Markus Funk (2018). "Quadcopter-projected in-situ navigation cues for improved location awareness". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–6.

Knorr, Alexander G, Lina Willacker, Joachim Hermsdörfer, Stefan Glasauer, and Melanie Krüger (2016). "Influence of person-and situation-specific characteristics on collision avoidance behavior in human locomotion." In: *Journal of experimental psychology: human perception and performance* 42.9, p. 1332.

Kollmitz, Marina, Andreas Eitel, Andres Vasquez, and Wolfram Burgard (2019). "Deep 3D perception of people and their mobility aids". In: *Robotics and Autonomous Systems* 114, pp. 29–40.

Krüger, Melanie, Rohan Puri, Jeffery J Summers, and Mark R Hinder (2024). "Influence of age and cognitive demand on motor decision making under uncertainty: a study on goal directed reaching movements". In: *Scientific Reports* 14.1, p. 9119.

Kumaran, Radha, You-Jin Kim, Anne E Milner, Tom Bullock, Barry Giesbrecht, and Tobias Höllerer (2023). "The Impact of Navigation Aids on Search Performance and Object Recall in Wide-Area Augmented Reality". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.

Lang, Alex H, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom (2019). "Pointpillars: Fast encoders for object detection from point clouds". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12697–12705.

Langbehn, Eike, Gerd Bruder, and Frank Steinicke (2016). "Subliminal reorientation and repositioning in virtual reality during eye blinks". In: *Proceedings of the 2016 symposium on spatial user interaction*, pp. 213–213.

Lappe, Markus, Frank Bremmer, and Albert V van den Berg (1999). "Perception of self-motion from visual flow". In: *Trends in cognitive sciences* 3.9, pp. 329–336.

Laureshyn, Aliaksei, Åse Svensson, and Christer Hydén (2010). "Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation". In: *Accident Analysis & Prevention* 42.6, pp. 1637–1646.

Lawson, Anna, Ieva Eskytė, Maria Orchard, Dick Houtzager, and Edwin Luitzen De Vos (2022). "Pedestrians with Disabilities and Town and City Streets: From Shared to Inclusive Space?" In: *The Journal of Public Space* 7.2, pp. 41–62.

Lees, Emily, Wendell C Taylor, Joseph T Hepworth, Karina Feliz, Andrea Cassells, and Jonathan N Tobin (2007). "Environmental changes to increase physical activity: perceptions of older urban ethnic-minority women". In: *Journal of aging and physical activity* 15.4, pp. 425–438.

Lehsing, Christian and Ilja T Feldstein (2018). "Urban interaction–getting vulnerable road users into driving simulation". In: *UR: BAN Human Factors in Traffic: Approaches for Safe, Efficient and Stress-free Urban Traffic*, pp. 347–362.

Li, Changyang, Wanwan Li, Haikun Huang, and Lap-Fai Yu (2022a). "Interactive augmented reality storytelling guided by scene semantics". In: *ACM Transactions on Graphics (TOG)* 41.4, pp. 1–15.

Li, Fei, Shiwei Fan, Pengzhen Chen, and Xiangxu Li (2020). "Pedestrian motion state estimation from 2D pose". In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1682–1687.

Li, Yao, Vinu Kamalasanan, Mariana Batista, and Monika Sester (2022b). "Improving pedestrians traffic priority via grouping and virtual lanes in shared spaces (short paper)". In: *15th International Conference on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Lindeman, Robert W, Robert Page, Yasuyuki Yanagida, and John L Sibert (2004). "Towards full-body haptic feedback: the design and deployment of a spatialized vibrotactile feedback system". In: *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 146–149.

Linder, Timm, Kilian Y Pfeiffer, Narunas Vaskevicius, Robert Schirmer, and Kai O Arras (2020). "Accurate detection and 3D localization of humans using a novel YOLO-based RGB-D fusion approach and synthetic training data". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1000–1006.

Ling, Haibin and Stefano Soatto (2007). "Proximity distribution kernels for geometric context in category recognition". In: *2007 IEEE 11th international conference on computer vision*. IEEE, pp. 1–8.

Linowes, Jonathan and Krystian Babilinski (2017). *Augmented reality for developers: Build practical augmented reality applications with unity, ARCore, ARKit, and Vuforia*. Packt Publishing Ltd.

Löcken, Andreas, Tamara von Sawitzky, Janine Bauer, and Andreas Riener (2023). "Exploring the Potential of eHMIs as Traffic Light Alternatives". In: *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 99–104.

Luiten, Jonathon, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe (2021). "Hota: A higher order metric for evaluating multi-object tracking". In: *International journal of computer vision* 129, pp. 548–578.

Lyssenko, Maria, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel (2021). "From evaluation to verification: Towards task-oriented relevance metrics for pedestrian detection in safety-critical domains". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38–45.

– (2022). "Towards safety-aware pedestrian detection in autonomous systems". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 293–300.

MacEachren, Alan M (2004). *How maps work: representation, visualization, and design*. Guilford Press.

Mao, Qi-Chao, Hong-Mei Sun, Yan-Bo Liu, and Rui-Sheng Jia (2019). "Mini-YOLOv3: real-time object detector for embedded applications". In: *Ieee Access* 7, pp. 133529–133538.

Markkula, Gustav, Ruth Madigan, Dimitris Nathanael, Evangelia Portouli, Yee Mun Lee, André Dietrich, Jac Billington, Anna Schieben, and Natasha Merat (2020). "Defining interactions: A conceptual framework for understanding interactive behaviour in human and automated road traffic". In: *Theoretical Issues in Ergonomics Science* 21.6, pp. 728–752.

Marquardt, Alexander, Jens Maiero, Ernst Kruijff, Christina Trepkowski, Andrea Schwandt, André Hinkenjann, Johannes Schöning, and Wolfgang Stuerzlinger (2018). "Tactile hand motion and pose guidance for 3d interaction". In: *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10.

Marszałek, Marcin, Cordelia Schmid, Hedi Harzallah, and Joost Van De Weijer (2007). "Learning object representations for visual object class recognition". In: *Visual Recognition Challange workshop, in conjunction with ICCV*.

Martin-Martin, Roberto, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese (2021). "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments". In: *IEEE transactions on pattern analysis and machine intelligence*.

Maruta, Kazuki, Miyuu Takizawa, Ryuichi Fukatsu, Yue Wang, Zongdian Li, and Kei Sakaguchi (2021). "Blind-spot visualization via AR glasses using millimeter-wave V2X for safe driving". In: *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*. IEEE, pp. 1–5.

Meerhoff, Laurentius Antonius, Julien Pettré, Sean Dean Lynch, Armel Crétual, and Anne-Hélène Olivier (2018). "Collision avoidance with multiple walkers: Sequential or simultaneous interactions?" In: *Frontiers in psychology* 9, p. 368231.

Merriam-Webster, Inc (1990). *Webster's ninth new collegiate dictionary*. Vol. 10. Merriam-Webster.

Microsoft (2023). *MRTK*. `https://learn.microsoft.com/en-us/windows/mixed-reality/`. Accessed: 01/08/2023.

Milgram, Paul, Haruo Takemura, Akira Utsumi, and Fumio Kishino (1995). "Augmented reality: A class of displays on the reality-virtuality continuum". In: *Telemanipulator and telepresence technologies*. Vol. 2351. Spie, pp. 282–292.

Miller, Ronald and Qingfeng Huang (2002). "An adaptive peer-to-peer collision warning system". In: *Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No. 02CH37367)*. Vol. 1. IEEE, pp. 317–321.

Monastero, Beatrice and David K McGookin (2018). "Traces: Studying a public reactive floor-projection of walking trajectories to support social awareness". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Morrison, Joel L (1977). "The science of cartography and its essential processes". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 14.1, pp. 58–71.

Moussaïd, Mehdi, Dirk Helbing, and Guy Theraulaz (2011). "How simple rules determine pedestrian behavior and crowd disasters". In: *Proceedings of the National Academy of Sciences* 108.17, pp. 6884–6888.

Mukbil, Awad, Yasin Yousif, Sakif Hossain, and Jörg P Müller (2023). "CTV-Dataset: A Shared Space Drone Dataset for Cyclist-Road User Interaction Derived from Campus Experiments". In: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 3186–3191.

Narzt, Wolfgang, Gustav Pomberger, Alois Ferscha, Dieter Kolb, Reiner Müller, Jan Wieghardt, Horst Hörtner, and Christopher Lindinger (2006). "Augmented reality navigation systems". In: *Universal Access in the Information Society* 4, pp. 177–187.

Nelson, Michael G, Alexandros Koilias, Dominic Kao, and Christos Mousas (2023). "Effects of Speed of a Collocated Virtual Walker and Proximity Toward a Static Virtual Character on Avoidance Movement Behavior". In: *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, pp. 930–939.

Neth, Christian T, Jan L Souman, David Engel, Uwe Kloos, Heinrich H Bulthoff, and Betty J Mohler (2012). "Velocity-dependent dynamic curvature gain for redirected walking". In: *IEEE transactions on visualization and computer graphics* 18.7, pp. 1041–1052.

Nguyen, Uyen Dao-Xuan (2020). "3D Pedestrian Tracking Using Neighbourhood Constraints". PhD thesis. Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover.

Nilsson, Niels Christian, Stefania Serafin, Frank Steinicke, and Rolf Nordahl (2018). "Natural walking in virtual reality: A review". In: *Computers in Entertainment (CIE)* 16.2, pp. 1–22.

Nyvig, Ramboll (2007). "Shared Space–Safe Space". In: *Meeting the Requirements of Blind and Partially Sighted People [Research report, commissioned by The Guide Dogs for the Blind Association]. Reading: the Guide Dogs for the Blind Association.*

Olaverri-Monreal, Cristina, Pedro Gomes, Ricardo Fernandes, Fausto Vieira, and Michel Ferreira (2010). "The See-Through System: A VANET-enabled assistant for overtaking maneuvers". In: *2010 IEEE intelligent vehicles symposium*. IEEE, pp. 123–128.

Olivier, Anne-Hélène, Julien Bruneau, Richard Kulpa, and Julien Pettré (2017). "Walking with virtual people: Evaluation of locomotion interfaces in dynamic environments". In: *IEEE transactions on visualization and computer graphics* 24.7, pp. 2251–2263.

Olivier, Anne-Hélène, Antoine Marin, Armel Crétual, Alain Berthoz, and Julien Pettré (2013). "Collision avoidance between two walkers: Role-dependent strategies". In: *Gait & posture* 38.4, pp. 751–756.

Olivier, Anne-Hélène, Jan Ondřej, Julien Pettré, Richard Kulpa, and Armel Cretual (2010). "Interaction between real and virtual humans during walking: Perceptual evluation of a simple device". In: *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pp. 117–124.

Orschiedt, Jonathan, Johanna Schmickler, Valentin Nußer, Timo Fischer, Joachim Hermsdörfer, and Melanie Krüger (2023). "Writing while walking: the impact of cognitive-motor multi-tasking on collision avoidance in human locomotion". In: *Human movement science* 88, p. 103064.

Orsini, Federico, Mariana Batista, Bernhard Friedrich, Massimiliano Gastaldi, and Riccardo Rossi (2023). "Before-after safety analysis of a shared space implementation". In: *Case Studies on Transport Policy*, p. 101021.

Papadopoulos, Theofilos, Konstantinos Evangelidis, Theodore H Kaskalis, Georgios Evangelidis, and Stella Sylaiou (2021). "Interactions in augmented and mixed reality: An overview". In: *Applied Sciences* 11.18, p. 8752.

Papageorgiou, George, Katerina Hadjigeorgiou, and Alexander N Ness (2020). "An innovative way to promote walking via a smartphone pedestrian navigation application". In: *2020 European Navigation Conference (ENC)*. IEEE, pp. 1–9.

Peck, Tabitha C, Henry Fuchs, and Mary C Whitton (2011). "An evaluation of navigational ability comparing redirected free exploration with distractors to walking-in-place and joystick locomotio interfaces". In: *2011 IEEE Virtual Reality Conference*. IEEE, pp. 55–62.

Peereboom, Joris, Wilbert Tabone, Dimitra Dodou, and Joost De Winter (2023). "Head-locked, world-locked, or conformal diminished-reality? An examination of different AR solutions for pedestrian safety in occluded scenarios". In: *ResearchGate*.

Peitso, Loren E and James Bret Michael (2020). "The promise of interactive shared augmented reality". In: *Computer* 53.1, pp. 45–52.

Philion, Jonah, Amlan Kar, and Sanja Fidler (2020). "Learning to evaluate perception models using planner-centric metrics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14055–14064.

Podkosova, Iana and Hannes Kaufmann (2018a). "Co-presence and proxemics in shared walkable virtual environments with mixed colocation". In: *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11.

– (2018b). "Mutual collision avoidance during walking in real and collaborative virtual environments". In: *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 1–9.

Qi, Charles R, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas (2018). "Frustum pointnets for 3d object detection from rgb-d data". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 918–927.

Qi, Charles R, Hao Su, Kaichun Mo, and Leonidas J Guibas (2017). "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.

Rameau, François, Hyowon Ha, Kyungdon Joo, Jinsoo Choi, Kibaek Park, and In So Kweon (2016). "A real-time augmented reality system to see-through cars". In: *IEEE transactions on visualization and computer graphics* 22.11, pp. 2395–2404.

Randhavane, Tanmay, Aniket Bera, Emily Kubin, Austin Wang, Kurt Gray, and Dinesh Manocha (2019). "Pedestrian dominance modeling for socially-aware robot navigation". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 5621–5628.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi (2016). "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Redmon, Joseph and Ali Farhadi (2018). "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767*.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28.

Roshan Zamir, Amir, Afshin Dehghan, and Mubarak Shah (2012). "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*. Springer, pp. 343–356.

Roth, Robert E (2017). "Visual variables". In: *International encyclopedia of geography: People, the earth, environment and technology*, pp. 1–11.

Rouchitsas, Alexandros and Håkan Alm (2019). "External human–machine interfaces for autonomous vehicle-to-pedestrian communication: A review of empirical work". In: *Frontiers in psychology* 10, p. 2757.

Sager, Christoph, Patrick Zschech, and Niklas Kühl (2021). "labelcloud: A lightweight domain-independent labeling tool for 3d object detection in point clouds". In: *arXiv preprint arXiv:2103.04970*.

Sakamoto, Nobuhito, Masahiro Furukawa, Masataka Kurokawa, and Taro Maeda (2019). "Guided Walking to Direct Pedestrians Toward the Same Destination". In: *Proceedings of the 10th Augmented Human International Conference 2019*, pp. 1–8.

Santos, Marc Ericson C, Takafumi Taketomi, Christian Sandor, Jarkko Polvi, Goshiro Yamamoto, and Hirokazu Kato (2014). "A usability scale for handheld augmented reality". In: *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, pp. 167–176.

Sauter, Daniel and Marco Huettenmoser (2008). "Liveable streets and social inclusion". In: *Urban Design International* 13.2, pp. 67–79.

Schall Jr, Mark C, Michelle L Rusch, John D Lee, Jeffrey D Dawson, Geb Thomas, Nazan Aksan, and Matthew Rizzo (2013). "Augmented reality cues and elderly driver hazard perception". In: *Human factors* 55.3, pp. 643–658.

Schroeder, Will, Kenneth M Martin, and William E Lorensen (1998). *The visualization toolkit an object-oriented approach to 3D graphics*. Prentice-Hall, Inc.

Seipel, Stefan (2013). "Evaluating 2D and 3D geovisualisations for basic spatial assessment". In: *Behaviour & Information Technology* 32.8, pp. 845–858.

Shaheen, Susan, Adam Cohen, Ismail Zohdy, and Beaudry Kock (2016). "Smartphone applications to influence travel choices: practices and policies". In.

Shenoi, Abhijeet, Mihir Patel, JunYoung Gwak, Patrick Goebel, Amir Sadeghian, Hamid Rezatofighi, Roberto Martin-Martin, and Silvio Savarese (2020). "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 10335–10342.

Shi, Guangsheng, Ruifeng Li, and Chao Ma (2022). "Pillarnet: Real-time and high-performance pillar-based 3d object detection". In: *European Conference on Computer Vision*. Springer, pp. 35–52.

Space, Shared (2005). "Shared space: Room for everyone". In: *Interreg IIIB Project 'Shared Space', Leeuwarden, The Netherlands*.

Stein, Niklas (2021). "Analyzing visual perception and predicting locomotion using virtual reality and eye tracking". In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, pp. 727–728.

Steinicke, Frank, Gerd Bruder, Jason Jerald, Harald Frenz, and Markus Lappe (2009). "Estimation of detection thresholds for redirected walking techniques". In: *IEEE transactions on visualization and computer graphics* 16.1, pp. 17–27.

Strauss, Ryan R, Raghuram Ramanujan, Andrew Becker, and Tabitha C Peck (2020). "A steering algorithm for redirected walking using reinforcement learning". In: *IEEE transactions on visualization and computer graphics* 26.5, pp. 1955–1963.

Suma, Evan A, Gerd Bruder, Frank Steinicke, David M Krum, and Mark Bolas (2012). "A taxonomy for deploying redirection techniques in immersive virtual environments". In: *2012 IEEE Virtual Reality Workshops (VRW)*. IEEE, pp. 43–46.

Sutcliffe, David (2009). "Shared Space and Naked Intersections". In.

Svensson, Åse and Christer Hydén (2006). "Estimating the severity of safety related behaviour". In: *Accident Analysis & Prevention* 38.2, pp. 379–385.

Tahara, Tomu, Takashi Seno, Gaku Narita, and Tomoya Ishikawa (2020). "Retargetable AR: Context-aware augmented reality in indoor scenes based on 3D scene graph". In: *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, pp. 249–255.

Takahashi, Masahiro, Yonghoon Ji, Kazunori Umeda, and Alessandro Moro (2020). "Expandable YOLO: 3D object detection from RGB-D images". In: *2020 21st International Conference on Research and Education in Mechatronics (REM)*. IEEE, pp. 1–5.

Tang, Liu and Jia Zhou (2020). "Usability assessment of augmented reality-based pedestrian navigation aid". In: *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Posture, Motion and Health: 11th International Conference, DHM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer, pp. 581–591.

Targ, Sasha, Diogo Almeida, and Kevin Lyman (2016). "Resnet in resnet: Generalizing residual architectures". In: *arXiv preprint arXiv:1603.08029*.

Thomas, Carol (2008). "Discussion: Shared space—safe space?" In: *Proceedings of the Institution of Civil Engineers-Municipal Engineer*. Vol. 161. 1. Thomas Telford Ltd, pp. 59–60.

Tonnis, Marcus, Christian Sandor, Gudrun Klinker, Christian Lange, and Heiner Bubb (2005). "Experimental evaluation of an augmented reality visualization for directing a car driver's attention". In: *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05)*. IEEE, pp. 56–59.

Toshev, Alexander and Christian Szegedy (2014). "Deeppose: Human pose estimation via deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660.

Uchiyama, Hajime, Michael A Covington, and Walter D Potter (2008). "Vibrotactile glove guidance for semi-autonomous wheelchair operations". In: *Proceedings of the 46th Annual Southeast Regional Conference on XX*, pp. 336–339.

Ungureanu, Dorin, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, et al. (2020). "Hololens 2 research mode as a tool for computer vision research". In: *arXiv preprint arXiv:2008.11239*.

Van Dyck, Delfien, Benedicte Deforche, Greet Cardon, and Ilse De Bourdeaudhuij (2009). "Neighbourhood walkability and its particular importance for adults with a preference for passive transport". In: *Health & place* 15.2, pp. 496–504.

Van Gemert, Jan C, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek (2009). "Visual word ambiguity". In: *IEEE transactions on pattern analysis and machine intelligence* 32.7, pp. 1271–1283.

Vanderbilt, Tom (2009). *Traffic: Why we drive the way we do (and what it says about us)*. Vintage.

Vassallo, Reid, Adam Rankin, Elvis CS Chen, and Terry M Peters (2017). "Hologram stability evaluation for Microsoft HoloLens". In: *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*. Vol. 10136. SPIE, pp. 295–300.

Villena-Martínez, Víctor, Andrés Fuster-Guilló, Jorge Azorín-López, Marcelo Saval-Calvo, Jeronimo Mora-Pascual, Jose Garcia-Rodriguez, and Alberto Garcia-Garcia (2017). "A quantitative comparison of calibration methods for RGB-D sensors using different technologies". In: *Sensors* 17.2, p. 243.

Wang, Jinbao, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao (2021). "Deep 3D human pose estimation: A review". In: *Computer Vision and Image Understanding* 210, p. 103225.

Wang, Leichen, Tianbai Chen, Carsten Anklam, and Bastian Goldluecke (2020). "High dimensional frustum pointnet for 3d object detection from camera, lidar, and radar". In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1621–1628.

Wang, Zhixin and Kui Jia (2019). "Frustum convnet: Sliding frustums to aggregate local pointwise features for amodal 3d object detection". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1742–1749.

Willett, Wesley, Yvonne Jansen, and Pierre Dragicevic (2016). "Embedded data representations". In: *IEEE transactions on visualization and computer graphics* 23.1, pp. 461–470.

Winkler, Susann, Juela Kazazi, and Mark Vollrath (2015). "Distractive or Supportive–How Warnings in the Head-up Display Affect Drivers' Gaze and Driving Behavior". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 1035–1040.

Wojke, Nicolai, Alex Bewley, and Dietrich Paulus (2017). "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE international conference on image processing (ICIP)*. IEEE, pp. 3645–3649.

Wolf, Mirja, Luiz R Douat, and Michael Erz (2021). "Safety-aware metric for people detection". In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, pp. 2759–2765.

Yang, Jackie, Christian Holz, Eyal Ofek, and Andrew D Wilson (2019). "Dreamwalker: Substituting real-world walking experiences with a virtual reality". In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 1093–1107.

Yoon, Ju Hong, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon (2016). "Online multi-object tracking via structural constraint event aggregation". In: *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 1392–1400.

Yu, Xinyan, Marius Hoggenmueller, and Martin Tomitsch (2023). "Your Way Or My Way: Improving Human-Robot Co-Navigation Through Robot Intent and Pedestrian Prediction Visualisations". In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 211–221.

Yuan, Yufei, Winnie Daamen, Bernat Goñi-Ros, and Serge P Hoogendoorn (2018). "Investigating cyclist interaction behavior through a controlled laboratory experiment". In: *Journal of transport and land use* 11.1, pp. 833–847.

Zeidler, Conrad, Matthias Klug, Gerrit Woeckner, Urte Clausen, and Johannes Schöning (2023). "ARCHIE 2: An Augmented Reality Interface with Plant Detection for Future Planetary Surface Greenhouses". In: *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, pp. 601–610.

Zhang, Weicheng, Hao Cheng, Fatema T Johora, and Monika Sester (2023). "ForceFormer: exploring social force and transformer for pedestrian trajectory prediction". In: *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1–7.

Zollmann, Stefanie, Tobias Langlotz, Raphael Grasset, Wei Hong Lo, Shohei Mori, and Holger Regenbrecht (2020). "Visualization techniques in augmented reality: A taxonomy, methods and patterns". In: *IEEE transactions on visualization and computer graphics* 27.9, pp. 3808–3825.

Acknowledgements

With these closing lines, i would like to thank all those people in having contributed towards the success of this thesis

I would like to thank my professor Dr.-Ing.habil. Monika Sester for all the support and faith in my efforts during the research. The insightful feedbacks shared by my supervisor have steered both my research and problem solving ability to a higher level. The freedom to independently carry out my research during my doctoral study phase helped me both to self organise and innovate with an interdisciplinary mindeset. I would also thank Prof. Dr-Ing. Claus Brenner for providing suggestions and comments that help me cross multiple hurdles in achieving my goals.

I am grateful to all the colleagues and professors from the DFG-funded research training group Socialcars. The colleagues from the group have deeply helped me to stay motivated and focused during the tough corona period and the difficult lock-downs. The diverse expertise of the group also helped me improve my research contributions via collaborations and joint efforts. A special mention to both Late Rahi Shet and Shule Li for always being around and filling our days with a smile.

While I thank all the other professors from the RTG Socialcars group for all their inputs, a special mention to my second supervisor, Prof Dr Jörg.P.Müller for both his technical guidance and professional advice during my complete journey.

A special thanks to Prof Melanie Krüger and all other collaborators who have both supported and help me in achieving the results that have been presented in this thesis. The experiments and publications completed jointly during the period has helped me achieve my goals, acquired the skills and expertise to complete my AR research.

A special mention to the colleagues from IKG for this involvement and advice during my research phase. The feedbacks and questions during the institute presentations (Monday Round) constantly helped me assess and improve my technical contributions. A special mention to Malte Schulze and Steffen Busch for hardware and support respectively in my work.

I would like to extend my sincere gratitude to professors and researchers from both the AR and geoinformatics community for the beautiful exchange of ideas and professional inputs about my work during the conference visits both virtually and in presence at the ISPRS, IEEE ISMAR, Eurographics and IEEE VR.

I would like to thank my wife Anupama, my little daughter Navomi and my family for having faith in me and helping me reach my research goals in a time bound manner.

Lastly i would like extend my sincere thanks to Deutscher Akademischer Austauschdienst (DAAD) for funding my research for the duration of my work and supporting me to stay in Germany to fulfil the requirements of this doctorate study.

Curriculum Vitae

Personal Information

Name	Vinu kamalasanan
Date of Birth	05 February 1989
Place of Birth	Dubai; United Arab Emirates
Nationality	Indian
Address	Vahrenwalder straße 59
	30165 Hannover, Germany

Education

06.2012-06.2014	**M.Tech. Communication Engineering**
	Vellore Institute of Technology, India
06.2007-06.2011	**B.Tech Electronics and Communication**
	Kerala University, India

Experience

Since 10.2019	**DAAD Research Scholar**
	Institute of Cartography and Geoinformatics, Leibniz University, Germany
09.2019-06.2016	**Senior Software Engineer**
	Robert Bosch, India
09.2016-06.2014	**Software Engineer**
	Robert Bosch, India