Computer Vision-Powered Applications for Interpreting and Interacting with Movement

Basel Nitham Hindi

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science
Thesis Track
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Computer Vision-Powered Applications for Interpreting and Interacting with Movement

Basel Nitham Hindi

Movement and our ability to perceive it are core elements of the human experience. To bridge the gap between artificial intelligence research and the daily lives of people, this thesis explores leveraging advancements in the field of computer vision to enhance human experiences related to movement. Through two projects, I leverage computer vision to aid Blind and Low Vision (BLV) people in perceiving sports gameplay, and provide navigation assistance for pedestrians in outdoor urban environments. I present Front Row, a system that enables BLV viewers to interpret tennis matches through immersive audio cues, along with StreetNav, a system that repurposes street cameras for real-time, precise outdoor navigation assistance and environmental awareness. User studies and technical evaluations demonstrate the potential of these systems in augmenting people's experiences perceiving and interacting with movement. This exploration also uncovers challenges in deploying such solutions along with opportunities in the design of future technologies.

Table of Contents

Acknowledgments x

Dedication xii

Chapter 1: Introduction 1

Chapter 2: Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts 4

2.1 Related Work 7

2.1.1 Approaches to Visual Media Accessibility 7

2.1.2 Sports Broadcast Accessibility 8

2.1.3 Sports Video Analysis via Computer Vision 9

2.2 Formative Study 9

2.2.1 Methods 9

2.2.2 Understanding BLV Viewers' Challenges of Watching Sports 10

2.2.3 Understanding BLV Viewers' Information Preferences for Watching Sports 11

2.2.4 Design Goals 12

2.3 Front Row: Immersive Audio Design 13

2.3.1 Visualizing Players' Positions 14

2.3.2 Visualizing Players' Shots 15

2.3.3 Blending Ambient Game Sounds 16

2.4 Front Row: Computer Vision Pipeline 17

2.4.1 Tracking the Court, Ball, and Players 17

2.4.2 Detecting Shots 19

2.4.3 Segmenting Rallies 20

2.5 Technical Evaluation 21

2.5.1 Procedure 21

2.5.2 Results 22

2.6 User Study 25

2.6.1 Study Description 26

2.6.2 Gameplay Understanding 28

2.6.3 Overall Experience 31

2.6.4 Forced Ranking Results 36

2.7 Discussion 37

2.8 Future Directions for Immersive Sports Audio 39

2.9 Applications 40

Chapter 3: StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians 43

3.1 Related Work 46

3.2 Formative Interviews 49

3.2.1 Methods 49

3.2.2 Findings: BLV Pedestrians’ Challenges in Outdoor Navigation 50

3.3 The StreetNav System 52

3.3.1 StreetNav: Technical Setup 52

3.3.2 StreetNav: Computer Vision Pipeline 53

3.3.3 StreetNav App: User Interface 58

3.4 User Study 63

3.4.1 Study Description 64

3.4.2 Results 68

3.4.3 Forced Ranking Results 76

3.4.4 How Individual Experiences Influenced Participants’ Preferences 77

3.5 Discussion 78

3.6 Limitations 81

Conclusion or Epilogue 83

References 85

List of Figures

2.1 A study participant, who is congenitally blind, using Front Row to watch a tennis match together with their sighted friend. Front Row is a system that automatically generates an immersive audio representation of a tennis broadcast video, allowing BLV viewers to more directly perceive what is happening in a tennis match. Front Row first recognizes gameplay from the video feed using computer vision, then renders players' positions and shots via spatialized (3D) audio cues. Front Row works with a standard pair of headphones. 5

2.2 Tennis gameplay as experienced (a) on TV via visuals by sighted viewers, and (b) on radio via announcer's description by BLV viewers. The visuals allow sighted viewers to perceive players' positions and actions to fully understand gameplay. The ball's path is indicated in yellow, and the players' movements are indicated in white. The radio descriptions, by contrast, convey a fraction of the information that visuals provide and do not offer a way to form one's own opinions of the game. 6

2.3 Front Row's 3D soundscape. The tennis court is displayed on a 2D plane orthogonal to the BLV viewer. Players' positions are represented by continuous humming sounds, and players' shots are represented by bell sounds similar to those in blind tennis [28]. These sounds are blended with the TV broadcast's original audio to incorporate ambient noises and the announcers' commentary. 14

2.4 Front Row's computer vision pipeline. The pipeline takes as input only the tennis broadcast video feed to generate audio representations of the game. It consists of three components: (a) tracking the court, ball, and players; (b) detecting shots: recognizing when, where, and how players hit a shot; and (c) segmenting rallies: identifying periods of play (as opposed to the many lull moments in between) to generate immersive audio only for these portions of the broadcast. 16

2.5 A sample video frame from a tennis TV broadcast showing (a) scenarios where court detection fails due to an audience member's white shirt and white advertisement boards, and (b) output from masking out the background using a segmentation model to mitigate these court detection failures. 18

2.6 Illustration of technique for detecting *when* a shot is hit. Change in the direction of the ball's trajectory within a fixed radial distance from either player is used to identify *when* shots are hit by a player. This change in direction is computed using player and ball coordinates with respect to the court's 2D representation. 18

2.7 Player tracking accuracy results. (a) The precision-recall curve at 0.5 IoU threshold. (b) The Confusion matrix at 0.5 IoU threshold and 0.5 confidence threshold. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model achieves a 97.2% mean average precision at 0.5 IoU threshold. 22

2.8 Confusion matrices for (a) detecting occurrence of shots and for (b) detecting the type of shots, i.e., forehands vs. backhands. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model correctly detects the occurrence of 80.8% shots, and classifies the shot types with 79.3% accuracy. 23

2.9 Participants' available choices for answering questions about (a) players' predominant shot types and (b) players' positions. We calculated participants' error rates by computing the distance between their response and the correct answer, which we illustrate here. A response of 'I don't know' was penalized more strongly, with a greater distance value. 29

2.10 Average distance errors for participants' responses to gameplay understanding across two metrics: (a) recognizing players' predominant shot types and (b) identifying players' positions. A Paired t-test revealed that Front Row was significantly ($p < 0.0001$) better than Radio, giving BLV participants an accurate understanding of the gameplay for both metrics. Error bars indicate standard error. 30

2.11 Average TLX scores for participants' overall experience. Participants rated their (a) perceived information overload, (b) perceived frustration, and (c) perceived immersion on a scale of 1—20 while viewing tennis rallies via the three audio formats. The error bars indicate standard error. Pairwise significance is depicted for $p < .01$ ($*$) and $p < .05$ ($**$). Participants rated Front Row as the most immersive tennis viewing experience of the three audio formats. 33

2.12 Forced ranking results. Participants ranked the three audio formats in order of their preference for viewing tennis games. Eight participants selected Front Row as their number one choice, four participants picked Radio as their top choice, while Television was unanimously ranked as the least preferred option by all participants. 35

2.13 Illustration of popular racket sports. Front Row's design for (a) tennis can be extended to make other racket sports, such as (b) badminton, (c) pickleball, and (d) table tennis, accessible to BLV viewers. 35

2.14 Front Row plug-in for video streaming platforms. Front Row can be integrated with online video streaming platforms, such as YouTube, ESPN+, and Hulu, to make recorded tennis broadcasts accessible to BLV viewers. This could work similarly to how closed captions are implemented on YouTube. Video source: Wimbledon's YouTube channel. 41

2.15 Recreational tennis game at a park. Front Row can make recreational tennis matches, such as matches at high schools, parks, and universities, accessible to BLV viewers. By processing a camera feed captured behind one of the players, Front Row can enable BLV audience members to follow the game in real time. Image source: The New York Times. 41

2.16 Gaming streams. A tennis video game stream by *Ray*, a popular streamer, who can be seen on the bottom right playing *Mario Tennis Aces*. Front Row can make both video game streams and video games themselves accessible to BLV viewers and gamers. Video source: Ray's YouTube channel. 42

3.1 StreetNav is a system that explores the concept of repurposing existing street cameras to support precise outdoor navigation for blind and low-vision (BLV) pedestrians. It comprises two components: (i) a computer vision (CV) pipeline, and (ii) a companion smartphone app. The computer vision pipeline processes the street camera's video feeds and delivers real-time navigation feedback via the app. StreetNav offers precise turn-by-turn directions to destinations while also providing real-time, scene-aware assistance to prevent users from veering off course, alert them of nearby obstacles, and facilitate safe street crossings. 44

3.2 Street camera used for StreetNav's development and evaluation. The camera (a) mounts on the building's second floor, and (b) captures the view of a four-way intersection. 53

3.3 Gesture-based localization for determining a user's position on the map. (a) A study participant (P1) is (c) prompted to wave one hand above their head, enabling the computer vision pipeline to distinguish them from other pedestrians in (b) the camera feed view and (d) the map. 55

3.4 StreetNav's internal graph representation for route planning. The user's current position is added dynamically as a start node to the graph upon choosing a destination. The shortest path, highlighted in green, is then calculated as per this graph representation. 56

3.5 Identifying obstacles in the user's vicinity. (a) A vehicle turning left yields to the BLV pedestrian (detected in purple) crossing the street. (b) StreetNav identifies the obstacles' category and relative location on the map to provide real-time feedback via the app. 57

3.6 Recognizing pedestrian signal states from the camera's video feed. StreetNav compares the number of white and red pixels in the signal crops to determine its state: (a) *walk* vs. (b) *wait*. 59

3.7 The StreetNav App's user interface. It provides routing instructions to their destination via (a) a path overview and (c, e) real-time feedback that updates their current instruction based on their location. Upon reaching a sidewalk, (b) the app informs the user about when it is safe to cross and (d) how much remains for them to cross over. It also (d) notifies the user of a nearby obstacle's category and relative location to help them avoid it. The app (f) confirms the user's arrival at the destination. Throughout the journey, the app provides (g) continuous audiohaptic feedback to prevent users from veering off track. 60

3.8 Audiohaptic cues for preventing users from veering off track. Sample user trajectories showing feedback when users veer (a) left, (b) do not veer, and (c) veer to the right. When the user's heading coincides with the route to the destination, within a tolerance angle θ (highlighted in green), users receive (b) subtle haptic vibrations to reinforce them. When they veer off the route, outside the tolerance angle θ, they hear spatialized beeping sounds that are rendered from the (a) right speaker when veering left, and from the (c) left speaker when veering right. 61

3.9 The routes used in the navigation tasks. (A) 12 meters, stationary person to avoid on the sidewalk. (B) 30 meters, cross street, and moving person to avoid on the sidewalk. (C) 38 meters, a 90° turn, cross street, and moving person to avoid on the crosswalk. To mitigate learning effects, routes for the two conditions are symmetrically designed, situated on opposite sides of the street. 65

3.10 Results for participants' experience with routing to the destination. Participants rated the (1) usefulness of routing instructions, and (2) the system's ability to track them en route to the destination. Participants found StreetNav'sturn-by-turn instructions significantly more useful and precise than BlindSquare's "as the crow flies"-style routing instructions. Pairwise significance is depicted for $p < 0.01$ ($*$) and $p < 0.05$ ($**$). The error bars indicate standard error. 69

3.11 Comparison of paths traveled by three participants (P1, P3, P8) for route 'C' using (a) StreetNav, and (b) BlindSquare. StreetNav's routing instructions consistently guided participants to the destination via the shortest path. BlindSquare, however, caused participants to take incorrect turns (P1, P3, P8), oscillate back and forth near destinations (P1, P8), and even go around the whole intersection before getting close to the destination (P8). 70

3.12 Results for participants' perceived ability to prevent veering off path. Participants rated their ability to (1) maintain a straight walking path, and (2) intuitiveness of the feedback regarding direction they should be moving in; on a scale of 1–5. StreetNav'saudiohaptic feedback was significantly more intuitive than BlindSquare's in preventing participants from veer off path. Pairwise significance is depicted for $p < 0.01$ (∗). The error bars indicate standard error. 71

3.13 Results for participants' perceived obstacle awareness. Participants rated their ability to (1) avoid obstacles, (2) identify its category (e.g., person, bicycle), and (3) determine its relative location; on a scale of 1–5. StreetNav significantly improved participants' awareness of nearby obstacles during navigation. Pairwise significance is depicted for $p < 0.01$ (∗) and $p < 0.05$ (∗∗). The error bars indicate standard error. 73

3.14 Results for participants' perceived comfort in crossing streets. Participants rated their perceived comfort in (1) making the decision on when to begin crossing the street, and in (2) pacing themselves when crossing. Participants were significantly more comfortable crossing streets with StreetNav in comparison to BlindSquare. Pairwise significance is depicted for $p < 0.01$ (∗) and $p < 0.05$ (∗∗). The error bars indicate standard error. 75

List of Tables

2.1 Self-reported demographics of our study participants. Five BLV participants (F1–F5) were recruited for the formative study (Section 2.2), while twelve BLV participants (P1–P12) were part of the user study evaluating Front Row (Section 2.6). Note that three participants from the formative study (F1-F3) also took part in the user study (P10-P12). Gender information was collected as a free response where our participants identified themselves as female (F), non-binary (NB), and male (M). The country codes refer to Bahrain (BH), India (IN), Saudi Arabia (SA), Singapore (SG), and the United States (US). Participants indicated their sports fandom as per Hunt et al.'s [93] scale which classifies sports fans into five categories: (1) temporary, (2) local, (3) devoted, (4) fanatical, and (5) dysfunctional. 24

3.1 Self-reported demographics of our participants. Gender information was collected as a free response; our participants identified themselves as female (F) or male (M). Participants rated their assistive technology (AT) familiarity on a scale of 1–5. . . . 50

3.2 Self-reported demographics of our study participants. Gender information was collected as a free response. Participants rated their familiarity with assistive technology (AT) on a scale of 1–5. 64

Acknowledgements

I would like to thank several people who have played pivotal roles in my journey. I am deeply grateful and very blessed to have had the support that I did throughout my graduate studies.

Thank you to my advisor Dr. Brian A. Smith for taking a bet on a summer researcher with no previous background in HCI, and then taking that same bet again to formalize me as a thesis student in his lab. The hours spent scribbling on Brian's whiteboard and learning how to frame research have been critical in my development as a researcher, writer, and presenter.

I am deeply grateful for my thesis committee members, Brian A. Smith, Tony Dear, and Brian Plancher for serving on the committee and for their feedback on my thesis.

Thank you to Mourad Elmalaoui for facilitating my transition towards computer science, sparking my graduate studies journey, and for advising me as one would advise a younger brother.

To my friends at Columbia's Computer-Enabled Abilities Laboratory, I hold our countless conversations and long nights in the highest regard. I extend particular thanks to Gaurav Jain who I worked closely with on both papers presented in this thesis, and who showed me the ropes of HCI and accessibility research.

I would like to acknowledge my co-authors on the two papers presented in this thesis. With regards to "Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts for Blind Viewers" (Chapter 2), I thank Gaurav Jain, Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, and Michael Malcolm, and Brian A. Smith. For "StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians" (Chapter 3), I thank Gaurav Jain, Zihao Zhang, Koushik Srinivasula, Mingyu Xie, Mahshid Ghasemi, Daniel Weiner,

Sophie Ana Paris, Xin Yi Therese Xu, Michael Malcolm, Mehmet Turkcan, Javad Ghaderi, Zoran Kostic, Gil Zussman, and Brian A. Smith.

Finally, I thank my family. Thank you to my mother, the embodiment of patience and selflessness. Thank you to my father, a critical and exemplary role model throughout many facets of my journey. Thank you to my siblings, who I wouldn’t trade for anyone in the world (with a few exceptions).

Dedication

This thesis is dedicated to the children of Palestine. For those who live through unrelenting injustice, for those whose lives have been forever altered by disability, and for those who knew death before knowing a day of freedom. After all, as a descendant of Palestinian refugees, had my grandparents made a different choice many years ago, I too would be in occupied Palestine facing this plight.

To the accessibility research community: our work is incomplete if we stand idly by as children are robbed of their sight, hearing, limbs, and mobility. Service of the disability community must include advocacy for peace and condemnation of violence and oppression everywhere. This demands courage and a willingness to use our voices, even when it may be uncomfortable to do so. Virtue is rarely convenient. If our voices protect a single life or avert one needless disability, then it will have been worth every effort.

To the children of Palestine, I ask that you forgive the failure of my words. Everything I say is less than what you deserve.

In solidarity for a free Palestine.

Chapter 1: Introduction

Movement is a central component of the human experience. Our interaction with movement and our perception of it shape our understanding of the world and influence how we interact with our environment. It plays a critical role in our cognitive development, social interaction, self awareness, spatial perception, and physical health [1, 2, 3]. Movement is as ubiquitous as it is impactful.

Sensors and computers, equipped with their capacity to capture and analyze dynamic data, have proven to be indispensable tools for gaining a greater understanding of movement. Previous human-computer interaction (HCI) research has harnessed technology to augment our experiences with movement. From intelligent prosthetics [4, 5] to wearable activity trackers [6, 7] and simulated motion in virtual reality [8, 9], movement is deeply embedded within HCI research. Recent strides in foundational computer vision (CV) models have unlocked new dimensions and pathways for this research [10, 11]. These breakthroughs have been diverse and impactful, from deriving physiological insights like heart-rate and blood-flow patterns from video, to tracking and controlling intricate robotic movements. As CV continues to evolve, a pressing need emerges – to consolidate these advancements into cohesive applications that center the human experience, through augmenting our ability to move and to perceive movement.

As a computer science researcher with a background in mechanical engineering and autonomous vehicles, I have always had a fascination with the intersection of technology and movement. My internship at Nike leveraging multi-modal AI for sports and web accessibility further solidified this interest. These experiences have directed my focus towards exploring the implications of AI innovations in domains closely tied to movement.

To close the gap between the advancements in CV and human movement experiences, this thesis embarks on two interconnected research projects. These projects serve as examples of harnessing the transformative capabilities of CV to understand and engage with movement. The first project revolves around helping users perceive motion from an external vantage point—a third-person perspective. Here, the aim is to enable Blind and Low Vision (BLV) users to interpret tennis gameplay through spatialized audio cues. The second project concerns interpreting and interacting with movement from a first-person perspective — from the heart of the action. In this project, CV is used to provide real-time navigation support for BLV pedestrians moving within urban landscapes.

In chapter 2, we present Front Row [12, 13], a system that automatically generates an immersive audio representation of sports broadcasts, specifically tennis, allowing BLV viewers to gain a better understanding of what is happening in the game. Front Row recognizes gameplay characteristics from the video feed using CV, then renders players' positions and shots via spatialized (3D) audio cues. User evaluations with 12 BLV participants show that Front Row gives BLV viewers a more accurate understanding of the game compared to TV and radio, enabling viewers to form their own opinions on players' movement and strategies. With this new understanding of BLV users' discernment of movement from a third-person view, we discuss future implications of Front Row and illustrate several applications.

In chapter 3, we repurpose street-embedded cameras to give BLV pedestrians real-time navigation assistance. Traditionally, BLV people rely on GPS-based systems for outdoor navigation. GPS's inaccuracy, however, causes them to veer off track, run into unexpected obstacles, and struggle to reach precise destinations. While prior work has made precise indoor navigation possible via additional hardware installations, enabling precise navigation outdoors remains a challenge. To address this, we present StreetNav [14, 15], a system that facilitates user interaction with street embedded cameras, such that BLV pedestrians can receive environmental awareness cues and navigation instructions in real-time. Within this work, we conduct user evaluations of StreetNav at the NSF PAWR COSMOS wireless edge-cloud testbed, showing that StreetNav guides people more precisely than GPS and provides richer environmental awareness. Through this exploration, we gain a new understanding of how CV can be used for interpreting and interacting with movement from a first-person perspective, and we discuss future implications for deploying such systems at scale.

In a period of technological development dominated by artificial intelligence, it is integral to consolidate technical achievements into functional HCI applications. Movement is a foundational part of the human experience, and through our two projects, this thesis explores CV's capability to augment and enhance this experience. It underscores the belief that purposeful and innovative technology can serve as a gateway to inclusivity, empowerment, and a connection with the dynamic movement of our world.

Chapter 2: Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts

Sports broadcasts are one of the most watched categories on TV for blind and low-vision (BLV) people, yet they remain inaccessible to BLV viewers [16, 17], making the experience of watching sports exclusionary and isolating for them [18, 19, 20]. BLV people use TV and radio to follow sports similar to sighted people, but find it difficult to fully understand what is happening in the game due to the lack of information conveyed via the broadcasts' audio. They must also rely on descriptions of the game from other people, such as sports commentators and friends they are watching with, to understand what is happening in the game. This means that others have the power to decide what BLV viewers should focus on, and that if others fail to describe a certain detail, there is no way for BLV people to access it. In short, BLV viewers have no way of visualizing exactly what is happening in sports broadcasts, and they are not afforded the agency to interpret what is happening for themselves.

Figure 2.2 shows the difference between sighted people's experience watching sports on TV and BLV people's experience following sports via descriptions more concretely. The TV visuals (Figure 2.2a) convey players' positions and actions thoroughly, allowing viewers to focus on the parts of the game they find interesting. The radio descriptions (Figure 2.2b), by contrast, are largely focused on Elena Rybakina, the far player in the TV broadcast. The announcer does not describe how Ons Jabeur runs across the court to the right to successfully play a shot, as seen in Figure 2.2a. Ultimately, we need to understand how to help BLV people more directly perceive sports broadcasts themselves instead of relying on others' descriptions.

In this work, we present *Front Row*, a system for automatically generating immersive audio representations of sports by inferring gameplay directly from a source broadcast video. The name

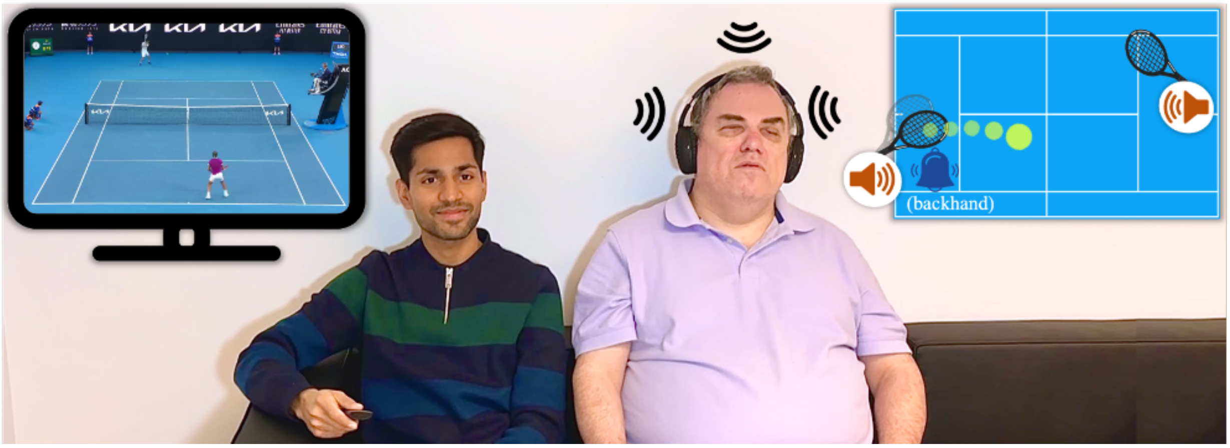
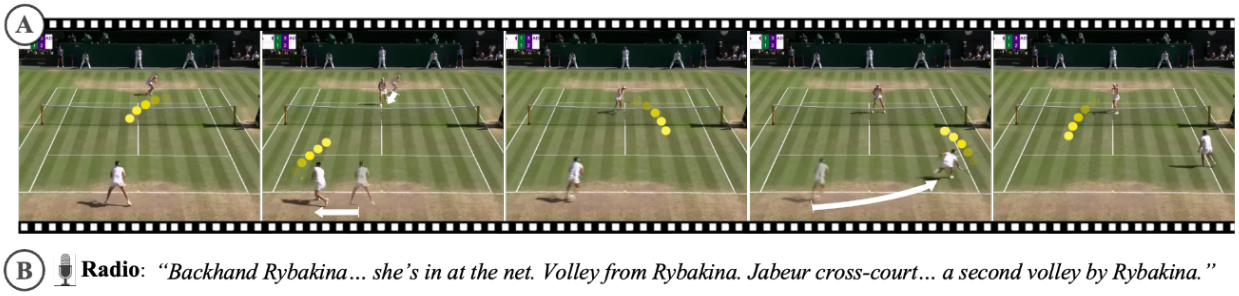


Figure 2.1: A study participant, who is congenitally blind, using Front Row to watch a tennis match together with their sighted friend. Front Row is a system that automatically generates an immersive audio representation of a tennis broadcast video, allowing BLV viewers to more directly perceive what is happening in a tennis match. Front Row first recognizes gameplay from the video feed using computer vision, then renders players' positions and shots via spatialized (3D) audio cues. Front Row works with a standard pair of headphones.

"Front Row" refers to our focus on giving BLV viewers a front row seat to the action so they can experience sports more immersively rather than relying on others' descriptions of the action. Front Row first uses a computer vision pipeline to automatically extract gameplay information from the broadcast video, then renders an immersive spatialized audio representation of the game to BLV viewers. The auto-generated spatialized audio cues convey players' positions and actions, enabling BLV viewers to visualize the action themselves. As Figure 2.1 illustrates, Front Row makes it possible for BLV people to enjoy sports together with friends without missing out on any important context.

Prior work has explored the use of on-field sensors such as high-precision cameras to generate audio and tactile representations of sports broadcasts [21, 22, 23, 24, 25, 26]. For example, Action Audio [21] acquires the ball's position using a specialized tracking system [27] that requires the court to be instrumented with multiple high-performance cameras. The use of specialized hardware, however, limits the applicability of these approaches to the tiny fraction of sports broadcasts where such large-scale hardware installations are feasible. With Front Row, we aim to use computer vision to generate immersive audio representations directly from the source broadcast video.



Radio: *"Backhand Rybakina... she's in at the net. Volley from Rybakina. Jabeur cross-court... a second volley by Rybakina."*

Figure 2.2: Tennis gameplay as experienced (a) on TV via visuals by sighted viewers, and (b) on radio via announcer's description by BLV viewers. The visuals allow sighted viewers to perceive players' positions and actions to fully understand gameplay. The ball's path is indicated in yellow, and the players' movements are indicated in white. The radio descriptions, by contrast, convey a fraction of the information that visuals provide and do not offer a way to form one's own opinions of the game.

By using this direct video-to-audio methodology, systems like Front Row could eventually make all sports broadcasts accessible to BLV viewers.

Our current focus with Front Row is on tennis broadcasts. We chose tennis because it is popular in many parts of the world, has a fairly simple setup with two players and a ball, and is very similar in form to other racket sports such as badminton, table tennis, and squash. As we discuss in Section 2.7, the results we find for Front Row could translate well to racket sports in general.

We evaluate Front Row in a user study with twelve BLV participants to understand how well Front Row allows BLV viewers to comprehend tennis gameplay compared to the status quo of listening to TV and radio broadcasts. We found that Front Row provides BLV viewers with a significantly more accurate understanding of the gameplay compared to TV and radio. For instance, Front Row reduced BLV participants' comprehension errors compared to TV by over 90% in recognizing the type of shots players hit and around 85% in identifying when players approach the net during the play. We also found that Front Row facilitates more immersion, with many participants valuing how Front Row affords them the ability to visualize the gameplay and to form their own opinions about the players' moods and strategies during the game. Our participants who play blind tennis [28] expressed their enthusiasm for using Front Row in the future to review their opponent's style of play before a game. We illustrate several future applications of Front Row, including a Front Row plug-in for video streaming platforms to make all sports videos across the

Web accessible and immersive for BLV people.

In summary, we contribute (1) a formative study of BLV people's challenges in watching sports, (2) the Front Row system for automatically generating immersive audio representations of sports from a source broadcast video, and (3) both a technical evaluation and a user experience evaluation of Front Row.

2.1 Related Work

Our work builds from the following three main threads of research. (i) approaches to visual media accessibility, (ii) sports broadcast accessibility, and (iii) sports video analysis via computer vision.

2.1.1 Approaches to Visual Media Accessibility

One common approach for making visual media accessible to BLV people is through *text-based descriptions*. For example, BLV people understand images via alternative text (also known as "alt-text") [29, 30] and videos via audio descriptions (AD) [31, 32]. Many researchers have studied ways to create effective descriptions for BLV people, by proposing methods and guidelines for authoring descriptions [33, 34, 35, 36, 37, 38, 39] as well as by introducing tools to support and automate the process [40, 41, 42, 43, 38, 44]. Prior work, however, shows that descriptions do not provide BLV people a spatial understanding of the visual content [45, 46, 47]. Spatial understanding of the visual content is crucial for interacting with rich visual media, such as for watching TV [48, 16, 32], exploring museums [49, 50, 47, 51], playing video games [52, 53], and engaging with social media [42, 45, 43, 54].

Another approach to visual media accessibility is using tactile graphics, which conveys spatial information via touch [55, 56]. Tactile graphics have been successfully used to understand the spatial layout of paintings [57], floor plans [58, 59], and more [60, 61, 62, 63, 64, 65, 66]. Prior work has also explored finger-worn devices [67, 68, 69] that allow BLV people to access printed text by moving their finger along the text for added spatial context. However, BLV users explore the tactile surfaces and use finger-worn devices through touch, which makes it less suitable for perceiving dynamic visual media such as videos. This becomes even more difficult for sports videos, given the fast-paced and dynamic nature of sports.

Audio, in the form of sonification or audio-cues, has also been explored for general image accessibility [70, 71], as well as for particular forms of images such as time series charts [72, 73]. However, limited work has been done to make videos [41], specifically sports videos, accessible via audio. In this work, we explore how spatialized audio can be used to make tennis videos accessible to BLV people, with the aim of giving them the ability to more directly visualize the gameplay.

2.1.2 Sports Broadcast Accessibility

Sports play an important role in enhancing people's social and cultural lives [18, 19, 20]. However, BLV people often experience sports in isolation because many existing sports broadcasts remain inaccessible to them [17]. Past research has explored different ways of making sports broadcasts accessible to BLV people, leveraging tactile graphic displays for football games [23, 24, 25] and 3D spatialized audio for tennis games [21, 22].

Most approaches, however, rely on specialized hardware which may not always be feasible. For example, Action Audio [21] requires the court to be equipped with the Hawk-Eye ball tracking technology [27], before it can make the game accessible to BLV people. Installing and maintaining these tracking technologies involves high costs which are only feasible for a tiny fraction of all sports events. Our preliminary work on Front Row [14], by contrast, introduced the concept of inferring gameplay directly from the source broadcast video feed using computer vision, eliminating the reliance on hardware installations. In this work, we perform both a technical evaluation and a user experience evaluation of Front Row.

2.1.3 Sports Video Analysis via Computer Vision

Research within the computer vision community has explored techniques to analyze sports videos by tracking game elements such as actions [74], balls [75], and players [76] and developing new applications using them [77, 78, 79, 80]. For instance, Voeikov et al. [80] introduced a deep learning-based system for automatic refereeing in table tennis games. Ghosh et al. [79] proposed a framework to infer players' statistics such as reaction time, speed, and movement for badminton.

Although this research is very promising, much of the focus has been outside of accessibility contexts and does not consider how computer vision systems can help users themselves watch and better perceive sports. Our work explores how sports video analysis can be used explicitly for accessibility. That is, we will first develop a computer vision system for computers to visualize sports (in our case, tennis), and we will then design an assistive interface so that BLV users can visualize sports.

2.2 Formative Study

To inform Front Row's design, we conducted semi-structured interviews and observation sessions with five BLV participants. Specifically, we focus on answering two questions:

Q1. What challenges do BLV people face when watching sports?

Q2. What are BLV viewers' information preferences for achieving a better understanding of the gameplay?

2.2.1 Methods

Participants

We recruited five BLV participants (three males, two females; aged 23–60) by posting to social media platforms. Table 3.2 summarises the participants' information (F1–F5). All interviews were conducted remotely via Zoom and lasted for about 60–75 minutes. Participants were compensated $25 for this IRB approved study.

Procedure

To answer the first question about BLV people's challenges of watching sports, we used a recent Critical Incident Technique (CIT) [81], in which we asked participants to recall and describe a recent time when they watched sports. We asked participants to describe their likes and dislikes about this experience, challenges they faced while viewing the game, and ways in which they navigated those challenges.

To answer the second question about BLV people's information preferences, we observed participants as they viewed tennis games via television (TV) and radio broadcasts. We shared our screen over Zoom and played several short clips from professional tennis matches for both TV and radio. After each clip, we asked participants to describe the gameplay and elaborate on aspects of gameplay they wanted to learn more about.

Interview Analysis

We first transcribed the interviews in full and then performed thematic analysis [82] involving two members of our research team. Each researcher independently reviewed the interview transcripts to generate an initial set of codes using NVivo [83]. Subsequently, both researchers collaborated with the two BLV co-authors to iterate on the codes and identify emerging themes for each research question.

For the first question, two challenges emerged: *(i)* feeling excluded when co-watching sports with friends, and *(ii)* inappropriate amount of information. For the second question, two information preferences emerged: *(i)* preference for spatial information, and *(ii)* preference for neutral, objective information about the gameplay.

2.2.2 Understanding BLV Viewers' Challenges of Watching Sports

We found two major challenges that BLV people face when watching sports.

Feeling excluded when co-watching sports with friends and family

Our participants noted that it is challenging for them to co-watch sports with friends and family because of mismatched preferences for the mediums through which they watch sports. BLV people prefer radio commentary, whereas their sighted friends and family prefer a visual medium such as TV. F2 mentioned that not being able to watch sports through a medium they could equally understand made them feel excluded: "*Well, I feel like I was kind of left out with the family conversation.*" F1 explained that feelings of exclusion are even more pronounced for sports because: *you don't like having what is supposed to be fun, make you feel excluded*".

This finding aligns with prior research showing that sports is a social activity for BLV people [18, 19, 20] and that the sense of shared excitement and affiliation is a big motivation for BLV people to watch sports [17].

Inappropriate amount of information

We observed that participants felt underwhelmed when watching tennis on TV and overwhelmed when watching tennis on radio. Participants noted that, unlike other sports, TV commentators in tennis are silent during the play. As a result, participants lose interest: "*I feel like there's a lot of stuff that I'm just not getting in, so I don't feel very immersed in it. And so my mind wanders*" (F1). On the contrary, radio announcers spoke too fast for them to be able to follow the game events, which made them feel frustrated sometimes.

2.2.3 Understanding BLV Viewers' Information Preferences for Watching Sports

We discovered BLV viewers' two major information preferences for better understanding gameplay.

Preference for spatial information about the gameplay

After listening to tennis clips for both TV and radio, participants expressed desire to more closely follow *where* the actions were happening on court: "*I never got a sense of where they were* *hitting it on the court. Because I know when you're really playing, if the player is up close to the net, then you try to hit it back in the far corner, you know, to make them have to run to make the play. I didn't get a sense whether that was happening or not.*" (F2).

Preference for neutral, objective information about the gameplay

Participants expressed their preference for fact-based reporting of information versus information interpreted from someone else's perspective. For instance, "*the announcers [often] color things from their home team's perspective*" (F4), and if a BLV viewer supports the other team, they "*probably wouldn't want [announcers'] opinions as much because I could form my own opinions*" (F4).

2.2.4 Design Goals

Based on our formative study findings, we set forth the following design goals for Front Row:

G1: Facilitating spatial understanding of the gameplay.

As noted that perceiving spatial aspects of gameplay are difficult in a non-visual format (Section 2.2.3), one of our goals to intuitively facilitate a spatial understanding of the gameplay for BLV people.

G2: Providing an appropriate amount of information to facilitate immersion.

Since immersion within the game is important to BLV viewers (Section 2.2.2), one of our aims is to ensure that an enhanced gameplay understanding is not achieved at the cost of immersion.

G3: Providing a single format that both BLV and sighted viewers can enjoy.

To instill a sense of affiliation in their sports watching experience (Section 2.2.2), one of our goals is to provide a single, universal format that BLV people can co-watch with friends and family.

G4: Supporting agency in gameplay understanding.

As mentioned in Section 2.2.3, it is important for BLV people to form their own opinion about the gameplay. Thus, one of our aims is to provide factual information that enables BLV people to view the game from their own perspective.

2.3 Front Row: Immersive Audio Design

Front Row is a system that generates an immersive audio representation of a tennis broadcast video in order to enable BLV viewers to more directly perceive what is happening in a tennis match. The audio rendering consists of three sound cues that together help BLV viewers to gain a spatial understanding of the gameplay (**G1**), to feel more immersed within the game (**G2**), to enable co-watching with sighted peers (**G3**), and to form their own opinions on players' strategies (**G4**).

The first sound cue allows viewers to visualize and follow players' positions on the court. The second sound cue allows viewers to understand players' shots, including *when* players make shots and whether those shots are forehands or backhands. The third sound cue is the ambient game sounds from the broadcast video, such as audience cheers and umpire's calls, that provide a more realistic viewing experience to BLV people.

Figure 2.3 shows how Front Row renders the sound cues to the viewer. Front Row renders the sound cues via spatialized (3D) audio on a 2D plane that represents the "birds-eye view" of the court. This 2D plane is orthogonal to the viewer but several feet in front of them in the 3D soundscape. To generate spatialized sound, we used the Steam Audio toolkit for Unity [84], which provides a built-in head-related transfer function (HRTF) [85]. Our design for Front Row resulted from several co-design sessions with our two BLV co-authors and consideration of prior work [21, 22]. In the following subsections, we describe each of Front Row's three sound cues.

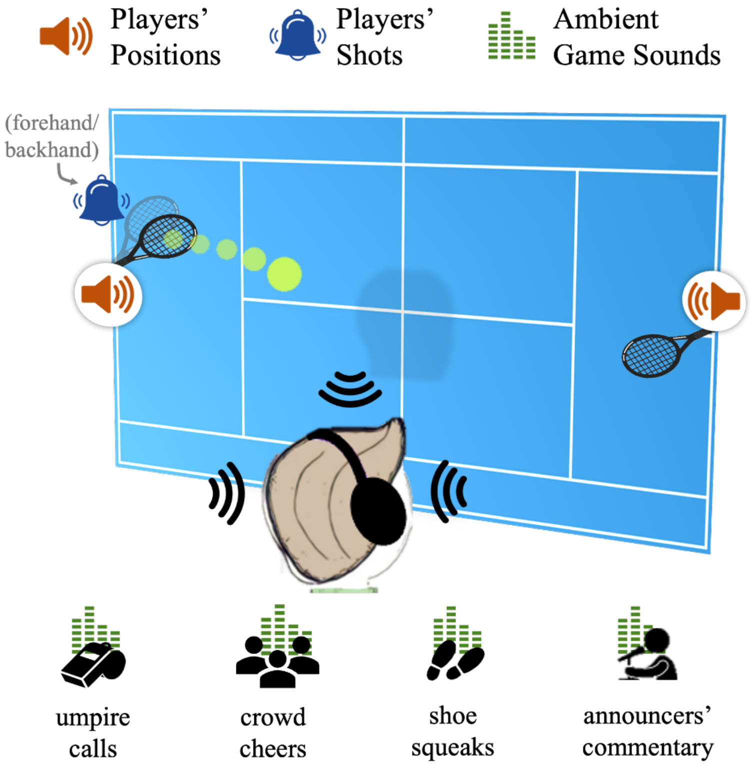


Figure 2.3: Front Row's 3D soundscape. The tennis court is displayed on a 2D plane orthogonal to the BLV viewer. Players' positions are represented by continuous humming sounds, and players' shots are represented by bell sounds similar to those in blind tennis [28]. These sounds are blended with the TV broadcast's original audio to incorporate ambient noises and the announcers' commentary.

2.3.1 Visualizing Players' Positions

Front Row renders each player's position via a virtual speaker that continuously emits a humming sound from the point on the 2D plane representing the player's position on the court. The humming sound uses a different pitch for each player. Effectively, viewers can hear virtual speakers moving in their left and right ears in sync with the player's movement on the court. Front Row renders the player shown closer in the TV broadcast on the left side and renders the player shown farther in the TV broadcast on the right side.

In our co-design sessions, we prototyped and evaluated different design possibilities for players' sounds, focusing on two main design "knobs": whether the sounds should be continuous or discontinuous (e.g., beeping or pulsing), and whether the 2D plane representing the court should be oriented orthogonally to the viewer (as we chose) or in a different fashion such as being parallel to the ground.

We compared a continuous sound effect with a discontinuous one because both are commonly used in blind-accessible video games to convey the position and movement of game objects [52, 53]. We experimented with a discontinuous sound representation where virtual speakers activate only when players pass by three points across the width of the court: the two ends and the middle. Both BLV co-authors agreed that while the discontinuous representation was less cognitively demanding, continuous representations provided a more immersive viewing experience (in line with **G2**). It allowed these co-authors to get a better feel for players' movements throughout the play without constantly needing to anticipate the players' positions during the gaps in the discontinuous representation (aligning with **G1**).

We tested different orientations of the court's 2D plane in order to see if a particular orientation made it easier for viewers to differentiate the two players and follow the action in general. We compared the court being parallel to ground, the court being orthogonal to the viewer but oriented horizontally, and the court being orthogonal to the viewer but oriented vertically.

Our BLV co-authors found it hard to clearly track the far player's movements when the court was rendered parallel to the ground. Comparing the two orthogonal representations, they found the horizontal configuration better at displaying continuous sounds since it allows viewers to hear each of the players' virtual speakers primarily in different ears. This aspect allows viewers to more easily alternate their focus to one side of the court as the ball moves around—a common practice for sighted viewers. Unlike the rendering scheme in Action Audio [21], which does not render players' positions and only uses discontinuous sounds via a vertical court orientation, Front Row's rendering scheme allows BLV viewers to continuously follow players' positions.

2.3.2 Visualizing Players' Shots

Front Row represents players' shots via differently pitched bell sounds that distinguish forehands from backhands. The bell sounds are rendered spatially from the player's location when

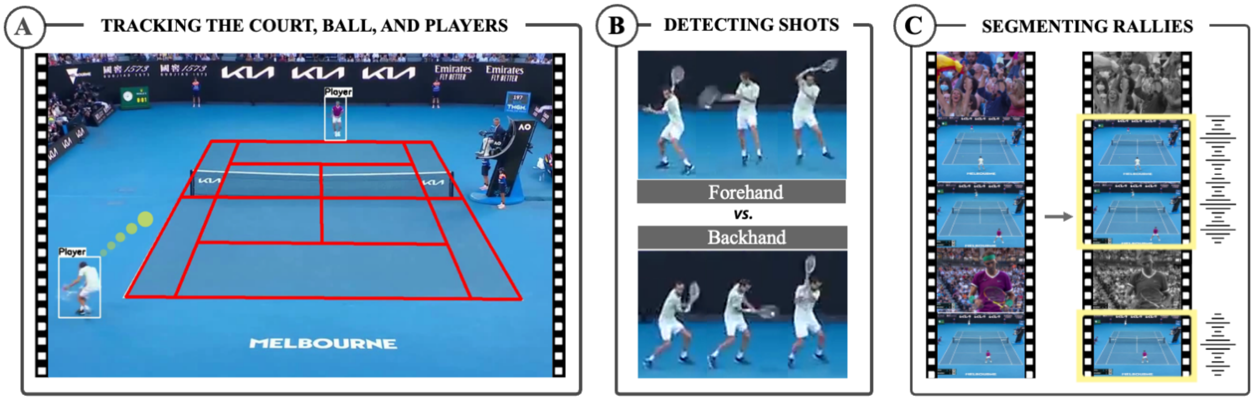


Figure 2.4: Front Row's computer vision pipeline. The pipeline takes as input only the tennis broadcast video feed to generate audio representations of the game. It consists of three components: (a) tracking the court, ball, and players; (b) detecting shots: recognizing when, where, and how players hit a shot; and (c) segmenting rallies: identifying periods of play (as opposed to the many lull moments in between) to generate immersive audio only for these portions of the broadcast.

they hit the shot. We chose a bell sound because it resembles the sound of the ball used in blind tennis [28], which was also the choice in prior work [21, 22]. Both BLV co-authors found it fairly easy to understand the ball's trajectory by interpolating the locations of two consecutive shots.

We had experimented with different ways of rendering shots as well. One interesting design that we prototyped was rendering the ball's position via a continuous sound cue, similar to the players' positions. Both BLV co-authors, however, found it extremely hard to follow a third sound cue that traveled back and forth between the left and right side. This led us to pursue a scheme for conveying the ball's trajectory indirectly via players' positions and shots.

2.3.3 Blending Ambient Game Sounds

To offer BLV viewers a more realistic and immersive viewing experience (**G2**), Front Row blends the audio from the source broadcast with the rest of the audio that it generates. We refer to the audio from the broadcast as *ambient game sounds*, which includes audio such as crowd cheers, the umpire's calls for faults and outs, sound from the rackets hitting the ball, players' grunts, TV announcers' commentary, and squeaking sounds caused by the friction between players' shoes and the court surface. These sounds can enhance viewers' comprehension of the gameplay (aligning with **G4**). Players' grunts, for instance, often indicate the intensity with which they hit a shot, while the squeaking sounds often give viewers a sense of the player's movements on the court. Note that the ambient sounds are rendered via mono audio since they do not correspond to a specific location in the 3D soundscape, unlike sound cues for players' positions and shots that are spatialized. Another reason that Front Row includes the ambient sounds is to afford a common context when BLV viewers watch the tennis match together with friends and family who are sighted. This way, all parties can hear the commentary from the broadcast, aligning with **G3**.

2.4 Front Row: Computer Vision Pipeline

Front Row's audio representations provide BLV viewers with information about players' positions and shots. To create these representations, Front Row takes as input only the source broadcast video feed and uses computer vision to extract the necessary gameplay information.

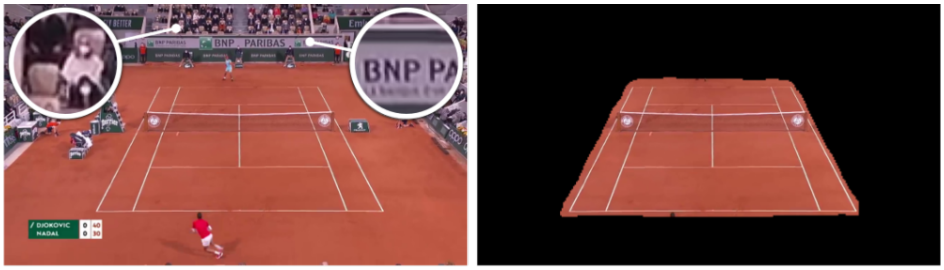
Figure 2.4 shows Front Row's computer vision pipeline. It consists of three components: (a) *tracking the court, ball, and players*; (b) *detecting shots*: recognizing *when*, *where*, and *how* players hit a shot; and (c) *segmenting rallies*: identifying periods of play (as opposed to the many lull moments in between). Front Row only generates immersive audio for the portions of the broadcast in which the ball is in play. The following subsections describe the computer vision pipeline's three components.

2.4.1 Tracking the Court, Ball, and Players

The first component in Front Row's computer vision pipeline tracks the basic game elements of tennis—the court, ball, and players—from the source broadcast video.

Tracking the Court

To track the court, we rely on the fact that court lines are always white in color. We first use thresholding to filter white pixels in the video feed image, and then we apply Hough Transforms [86] to identify white lines in the filtered image. From these candidate white lines, we select lines that match the expected structure of a tennis court, using perspective homography to find the closest match.



(a) (b)

Figure 2.5: A sample video frame from a tennis TV broadcast showing (a) scenarios where court detection fails due to an audience member's white shirt and white advertisement boards, and (b) output from masking out the background using a segmentation model to mitigate these court detection failures.

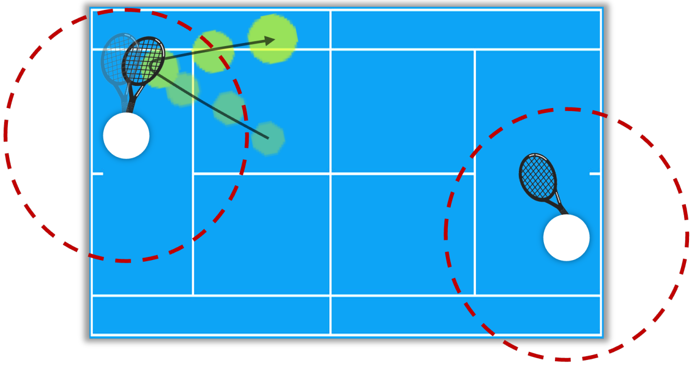


Figure 2.6: Illustration of technique for detecting *when* a shot is hit. Change in the direction of the ball's trajectory within a fixed radial distance from either player is used to identify *when* shots are hit by a player. This change in direction is computed using player and ball coordinates with respect to the court's 2D representation.

This approach correctly detects the court most of the time, but it sometimes confuses other white lines in the video feed as court lines. For example, Figure 2.5a shows a specific frame from a tennis match where we noticed failures in court detection due to a white advertisement board and audience members wearing white shirts. To address this problem, we compute a rough mask of the court area by using a semantic segmentation model [87], masking out the background as Figure 2.5b shows. We then detect white lines in the roughly masked court area only. This fix eliminated false detections of white lines completely.

Finally, we establish a court reference frame by transforming the detected court onto a reference court image. The reference court image is a “birds-eye view” of an actual tennis court. Recall that Front Row uses this court reference frame to establish the 3D soundscape, as seen in Figure 2.3.

Tracking the Ball

To track the ball, we used the state-of-the-art deep learning approach for detecting small, fast-moving objects — namely, TrackNet [75]. TrackNet outputs the ball’s pixel coordinates at every frame. We convert these pixel coordinates to “court coordinates” using the tracked court as a reference frame. In Section 2.4.2, we describe how the ball tracking is used to detect *when* a shot is played.

Tracking the Players

To track the players, we employ the YOLOv5 object detection model [88] to find the players’ positions in terms of pixel coordinates. We chose YOLOv5 [88] since it offers accurate detections at real-time speeds. The publically available pre-trained model, however, only has a ‘person’ class and not a specific ‘tennis player’ class, which means that it detects ball kids and line judges on and around the court as well. Thus, we annotated our own dataset and fine-tuned YOLOv5 using this dataset to accurately detect the two players. Our dataset features two classes: “Far player” and “Near player,” where “Far player” corresponds to the player farther away in the broadcast video feed. Now that we have pixel coordinates for both players, we convert them to “court coordinates” using the tracked court as a reference frame.

2.4.2 Detecting Shots

To help BLV viewers infer the shots hit by each player, Front Row’s audio representations need information about *when* a shot is played, *how* it is played, and *where* on the court it is played. Therefore, our computer vision pipeline should extract these three pieces of information about the players’ shots from the broadcast video feed.

To detect shots, we rely on the fact that the ball changes its direction perpendicular to the net whenever a player hits a shot. Since players hit shots when the ball is close to them, we only need to consider the ball's changes in direction when it happens near one of the players. Therefore, we use ball tracking and player tracking to identify moments when the ball changes direction. Figure 2.6 illustrates this technique, where we classify the ball's direction change as a shot when it happens within a fixed radial distance from the nearest player's position on the court. We determined the fixed radial distance empirically to optimize accuracy.

Now that we know *when* a shot is hit, we detect the specific shot type (forehand vs. backhand) using a recurrent neural network (LSTM [89, 90]). The recurrent network takes as input a sequence of 9 player crops and classifies the shot type. We select player crops from 9 consecutive video frames such that the middle frame corresponds to the moment ball changes direction in the player's vicinity. Our choice of 9 frames is based on empirical analysis of the average time taken by players to hit shots. Finally, we use the player's position as a proxy for *where* the shot is hit on the court.

2.4.3 Segmenting Rallies

With the components from Section 2.4.1 and Section 2.4.2, the computer vision pipeline has the ability to infer players' positions and players' shots from the broadcast video feed. Sports broadcasts, however, include a sequence of periods of play with lull periods interweaved within the game, where no action is happening. In tennis broadcasts, the play consists of rallies with non-play periods between them, such as commercial breaks, audience reactions, and players switching sides. A *rally* in tennis is analogous to what one might call a play or a point in other sports: it is an exchange of shots between players, ending when one player fails to make a successful return. Front Row renders the audio representations only for rallies in the tennis match. Thus, our computer vision pipeline should also segment the source broadcast video feed into rallies (Figure 2.4C).

To segment rallies from the broadcast video feed, we used the observation that during a rally, the camera is steadily positioned behind one of the players overlooking the full court. When a rally is not being played, the broadcast usually shows player or crowd close-ups. It may also be playing advertisements during breaks in game. Thus, to detect rallies, we trained a support vector classifier (SVC) [91] that detects the full view of the court in broadcast videos. The SVC takes as input the histogram of oriented gradients (HOG) [92] features extracted from each video frame and classifies the frame as a rally or non-rally frame. Front Row generates audio representations for only these rally segments, as shown in Figure 2.4C. The ambient game sounds, however, can be heard at all times during the game, even when the ball is not in play.

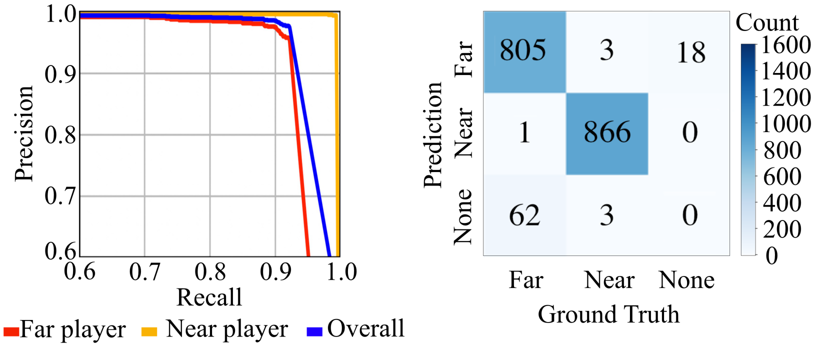
2.5 Technical Evaluation

We evaluate Front Row's technical performance to investigate the effect of errors on BLV viewers' experience of watching tennis via Front Rows' audio representation. We aim to answer two questions through this evaluation: (1) *To what extent does Front Row generate accurate audio representations of the game, and where does it fall short?* and (2) *How do the errors in Front Row affect BLV viewers' understanding of the game, and what strategies do BLV viewers use to compensate for system errors?*

2.5.1 Procedure

To answer the first question, we evaluate Front Row's ability to accurately convey the three main pieces of information it uses to render the audio representations, (i) players' positions: location of the humming sounds on the court, (ii) the occurrence of shots: *when* to play the bell sound cue, and (iii) type of shot: varying pitch of the bell sound to distinguish forehands from backhands. We perform the evaluation on a dataset of three videos of extended highlights from professional tennis broadcasts downloaded from YouTube. Each video is around five–six minutes long. To evaluate the pipeline's robustness to the court's visual appearance, we chose videos such that each tennis match was played on a different court surface. Thus, the matches corresponded to the three court surfaces in tennis, (i) synthetic: the blue court in Figure 2.1, (ii) grass: the green court in Figure 2.2, and (iii) clay: the red court in Figure 2.5.

To answer the second question, we conduct a pilot study with two BLV participants and gather initial reactions to Front Row's errors before performing the user study described later in Section 2.6. We recruited two additional BLV participants for this pilot study to ensure they had not tried Front Row before and were independent of our formative and user study participants. In the pilot study, we compare participants' experience watching tennis via Front Row in two conditions. The first corresponds to Front Row's actual accuracy performance, and the second corresponds to a version of Front Row with perfect accuracy performance. To prepare the perfect version, we manually corrected any errors in Front Row's computer vision pipeline before rendering the audio representations. We showed participants five tennis rally clips for each condition without revealing the condition name. After watching the clips, we asked participants questions to elicit differences in experiences between the two conditions.



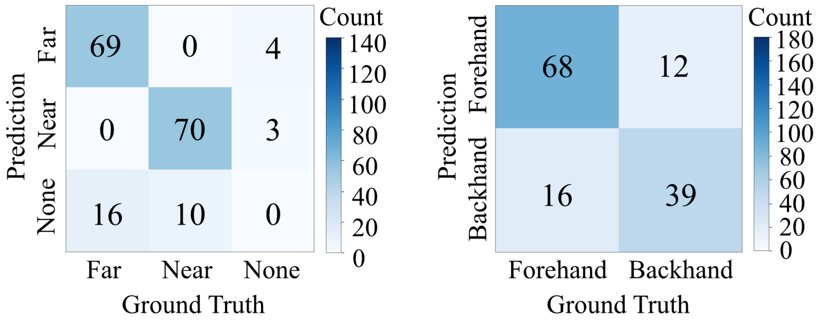
(a)

(b)

Figure 2.7: Player tracking accuracy results. (a) The precision-recall curve at 0.5 IoU threshold. (b) The Confusion matrix at 0.5 IoU threshold and 0.5 confidence threshold. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model achieves a 97.2% mean average precision at 0.5 IoU threshold.

2.5.2 Results

We present the results for our first question by reporting player tracking and shot detection performance.



(a) (b)

Figure 2.8: Confusion matrices for (a) detecting occurrence of shots and for (b) detecting the type of shots, i.e., forehands vs. backhands. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model correctly detects the occurrence of 80.8% shots, and classifies the shot types with 79.3% accuracy.

Player Tracking Accuracy

Figure 2.7 summarizes the accuracy with which Front Row tracks players' positions via a precision-recall curve (Figure 2.7a) and confusion matrix (Figure 2.7b). Our custom-trained player detection model scored a 97.2% mean average precision (mAP) at 0.5 intersection over union (IoU) threshold. We observed a minor accuracy drop when tracking the far player. Upon further analysis of the failure cases and the confusion matrix, we found that our pipeline sometimes confuses the ball kids in the background as the player. Another reason for the accuracy drop is the far players' size compared to the near player. A lower pixel resolution of the far player affects model performance.

Shot Occurrence Detection Accuracy

Figure 2.8a shows the confusion matrix for detecting the occurrence of shots. Our pipeline correctly detects 80.8% of the total shots. We noticed comparable performance for both players. Further analysis of failure cases revealed that most errors were attributed to the errors in our ball tracking approach, which uses TrackNet [75]. Future improvements in ball tracking could potentially increase shot detection accuracy.

Table 2.1: Self-reported demographics of our study participants. Five BLV participants (F1–F5) were recruited for the formative study (Section 2.2), while twelve BLV participants (P1–P12) were part of the user study evaluating Front Row (Section 2.6). Note that three participants from the formative study (F1-F3) also took part in the user study (P10-P12). Gender information was collected as a free response where our participants identified themselves as female (F), non-binary (NB), and male (M). The country codes refer to Bahrain (BH), India (IN), Saudi Arabia (SA), Singapore (SG), and the United States (US). Participants indicated their sports fandom as per Hunt et al.'s [93] scale which classifies sports fans into five categories: (1) temporary, (2) local, (3) devoted, (4) fanatical, and (5) dysfunctional.

PID	Gender	Age	Race	Country	Occupation	Vision ability	Onset	Sports Fandom (1–5)	Tennis Familiarity (1–5)
P1	M	27	Asian	IN	PhD student	Totally blind	Birth	4: Fanatical fan	1: Not at all familiar
P2	M	27	Arab	BH	Salesforce Admin	Totally blind	Birth	2: Local fan	3: Moderately familiar
P3	M	26	White	US	Student	Totally blind	Birth	1: Temporary fan	1: Not at all familiar
P4	M	23	Arab	SA	Student	Totally blind	Birth	5: Dysfunctional fan	2: Slightly familiar
P5	M	25	Asian	US	Not employed	Totally blind	Birth	3: Devoted fan	4: Very familiar
P6	F	52	Asian	SG	Massage therapist	Totally blind	Age 28	1: Temporary fan	2: Slightly familiar
P7	NB	40	White	US	Not employed	Low vision	Birth	1: Temporary fan	1: Not at all familiar
P8	F	25	Black	US	Not employed	Low vision	Age 10	4: Fanatical fan	5: Extremely familiar
P9	M	23	Latino	US	Customer service	Totally blind	Birth	1: Temporary fan	2: Slightly familiar
F1/P10	M	37	White	US	Game developer	Totally blind	Birth	4: Fanatical fan	5: Extremely familiar
F2/P11	F	60	White	US	Retired	Totally blind	Age 25	2: Local fan	2: Slightly familiar
F3/P12	F	28	White	US	FMLA claims expert	Totally blind	Birth	1: Temporary fan	2: Slightly familiar
F4	M	32	Asian	IN	Self-employed	Low vision	Age 20	1: Temporary fan	1: Not at all familiar
F5	M	23	Black	US	Editor	Low vision	Age 15	4: Fanatical fan	2: Slightly familiar

Shot Type Detection Accuracy

Figure 2.8b shows the confusion matrix for shot type detection accuracy. Our pipeline detected 79.3% of the shot types correctly. Our analysis of the failure cases revealed that the model struggled to correctly detect shot types for players that were left-handed or had unconventional ways of playing backhands. For example, most players use one hand for playing forehands and both for playing backhands. However, few players play a single-handed backhand which is not well represented in our training dataset. Training the model with a larger, more diverse dataset could potentially improve shot type detection performance.

Next, we present the results for our second question by reporting findings from our experiments with the two BLV participants.

Pilot Study Results

The majority of Front Row's errors were due to the computer vision pipeline's inability to accurately detect shots and their types. As a result, Front Row sometimes missed out on rendering the bell sound cue for a player's shot, or misrepresented the shot type (for e.g., representing a forehand as a backhand). While trying Front Row in the two conditions —the actual and the perfect version (with all errors removed)— both participants noticed these errors but remarked that they did not significantly affect their overall experience of viewing the game.

When Front Row fails to render the bell sound cue for a player's shot, the user loses information about the occurrence of a shot, its location on court, and the type. Both pilot study participants mentioned leveraging the two other sound cues in Front Row to recover a part of this lost information. To recognize the occurrence of a shot, both participants mentioned using Front Row's ambient game sounds (Section 2.3.3) which includes the sound of racket hitting the ball.

To identify the shot's location, one participant mentioned relying on the players' position sound cues (Section 2.3.1) at the time of the shot to get a general sense of the shot's location on court. The other participant remarked that since shots alternate between the two players, knowing which player hit the previous shot and knowing the occurrence of a shot via ambient game sounds was enough to keep them engaged within the game. While participants had no way of ascertaining the type of shot when Front Row failed to render it accurately, both participants agreed that this issue was not too common and thus, did not affect their understanding of the gameplay as much.

2.6 User Study

Our study had three goals. First, we wanted to evaluate how Front Row affects BLV viewers' ability to understand tennis gameplay compared to their existing means of viewing tennis: Television broadcasts and Radio broadcasts (Section 2.6.2). Second, we wanted to quantitatively analyse BLV people's overall experience of viewing tennis games using Front Row and these existing means (Section 2.6.3). Third, we wanted to see how participants rank the three audio formats

(Television, Radio, and Front Row) in order of their preference for viewing tennis games (Section 2.6.4).

2.6.1 Study Description

Participants

We recruited twelve BLV participants (seven males, four females, and one non-binary; aged 23–60) by posting to social media platforms and by snowball sampling [94]. Participants identified themselves with a range of racial identities (Asian, Black, White, Latino, Arab) and lived in five different countries (Bahrain, India, Saudi Arabia, Singapore, US). Participants also had diverse visual abilities, onset of vision impairment, sports fandom [93], and familiarity with tennis rules.

Table 3.2 summarises participants' information. P1 and P5 reported minor hearing loss in their right and left ear, respectively. All but three participants (P7, P9, and P10) reported themselves as being moderately–extremely experienced with 3D spatialized audio in the past (3+ scores on a 5-point Likert scale).

Experimental Design

In the study, participants had to answer questions about tennis audio clips in three formats: Television, Radio, and Front Row. The questions helped us quantify participants' understanding of the gameplay and their overall experience of viewing tennis games using each audio format.

Our study was a within-subjects design in which participants tried the three formats in a counter-balanced order. We used a balanced Latin square to counter-balance the order to reduce order bias and learning effects. For each audio format, participants listened to five audio clips rendered in that audio format. We gathered these clips from a single set of five rallies from different professional tennis matches. The length of each rally (and clip) was roughly ten to fifteen seconds. We extracted the Television and Radio audio clips from their official broadcasts, which we downloaded from YouTube. We generated the Front Row audio clips using Television broadcast video as input to our pipeline.

Procedure

We began each study condition (audio format) by playing a sample audio clip to help participants familiarize themselves with the format. For Front Row, we additionally gave a brief explanation about how to interpret its different audio cues. Participants were asked to wear a regular pair of headphones during the study to ensure optimal rendering of Front Row's spatialized audio.

We administered a post-clip questionnaire after each audio clip (3 audio formats × 5 rallies = 15 audio clips), which was comprised of three parts. The first part determined participants subjective understanding of the gameplay. It asked them to describe the gameplay in the rally. The second part tested participants' objective understanding of the gameplay. It included questions about players' predominant shot types and their positions. The third part gauged participants' overall experience via subjective measures of information overload, frustration, and immersion for the clip using 20-point Likert scales similar to a NASA TLX form [95]. We chose the objective measures for gameplay understanding and the subjective measure of participants' overall experience based on our formative study findings (Section 2.2).

After trying all three audio formats, participants completed a post-study questionnaire which asked them to rank the three audio formats in order of their preference for viewing tennis games. Last, we conducted a semi-structured interview to follow up on their responses to the questionnaires. Towards the end of the interview, we focused our discussion on Front Row, asking participants about ways in which it can be improved and scenarios in which they imagine themselves using Front Row.

The study was held virtually via Zoom and lasted for about 90–120 minutes. We ran studies at very different times of day to accommodate our participants' wide range of geographic locations. To play audio clips for participants over Zoom, the facilitator shared their screen's audio. Participants were compensated with a $25 gift card for their time. The study was IRB approved.

Interview Analysis

We report participants' spontaneous comments that best represent their overall opinions, providing further context on the quantitative data we collected during the study. We analyzed the transcripts for participants' quotes and grouped them according to (1) gameplay understanding, (2) overall experience, and (3) ranking preferences; across the three audio formats.

2.6.2 Gameplay Understanding

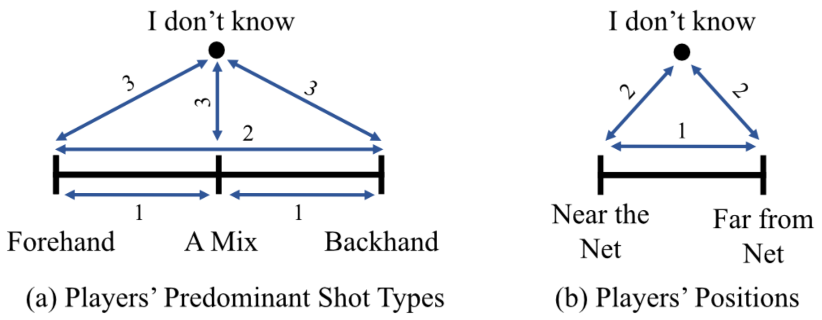
Here we report participants' ability to understand the gameplay using each audio format. We evaluate participants' gameplay understanding by computing participants' error in answering questions about two basic aspects of the game: *(i)* recognizing players' predominant shot types: what each player was doing, and *(ii)* identifying players' positions: where each player was on the court.

In the following subsections, we describe how we computed participants' errors, then compare participants' errors for the three audio formats — Television, Radio, and Front Row. We also elaborate on how participants' descriptions of the gameplay they viewed differed across the three formats.

How We Computed Participants' Errors

Figure 2.9 shows participants' available choices for these two questions and how we scored participants' responses. As Figure 2.9a shows, participants had to specify each player's predominant shot type for each rally from three choices: mostly forehands, mostly backhands, or a mix of the two. As Figure 2.9b shows, participants had to specify each player's position using two options: near the net and far from the net. For both questions, we gave participants the option to choose 'I don't know' if they had no idea at all.

We computed participants' error rates by calculating the distance of their response from the correct answer on the relevant spectrum from Figure 2.9. Note that we penalized 'I don't know' more strongly — with a greater distance value — since it reflected them not being able to ascertain any information at all.



(a) Players' Predominant Shot Types (b) Players' Positions

Figure 2.9: Participants' available choices for answering questions about (a) players' predominant shot types and (b) players' positions. We calculated participants' error rates by computing the distance between their response and the correct answer, which we illustrate here. A response of 'I don't know' was penalized more strongly, with a greater distance value.

Recognizing Players' Predominant Shot Types

Figure 2.10a shows the average error in participants' understanding of the shot types that players predominantly played. The mean (± std. dev.) error for Front Row was the least of the three conditions, at 0.21 (±0.22), followed by Radio in a distant second at 1.48 (±0.74) and Television last at 2.68 (±0.82). The error results for Television failed the Kolmogorov-Smirnov test for normality, i.e., it varied significantly from a normal distribution. Thus, we did not run any parametric tests on the Television error results. A paired t-test was performed on the error results for Radio and Front Row. Average error in participants' responses to shot types with Front Row were significantly ($t_{11} = 5.66, p < 0.0001$) lower than those with Radio. This indicates that Front Row gave BLV participants a more accurate understanding of the shot types compared to Radio.

Regarding participants' ability to describe the gameplay that they viewed, participants could only specify the number of shots in a rally when viewing the Television clips. Radio gave participants a general sense of what happened in the rally — which is more detail than simply the number of shots that occurred — but participants were still confused about the specifics of what happened when viewing the Radio clips:

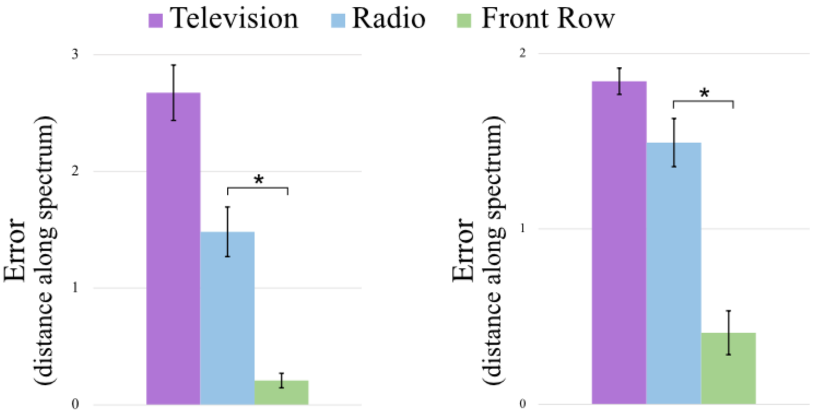
> P3: *"I understand that there's a couple backhands and then there's a forehand, but I don't know who's doing what. ... [The announcer] was mostly talking about like every third shot, so it's confusing."*

Front Row, on the contrary, enabled participants to follow the game more closely with access to information about almost every shot:

> P12: *'Well, that was eventful. Um, [it had] mix of forehands and backhands on both sides, and it culminated with a forehand from the player on the right."*

With Front Row, participants also liked how intuitively they could relate the shot types with specific players — something they mentioned missing with Radio:

> P11: *"I really liked the different pitches of the different shots. I [also] liked hearing shots on the left or right side of my headset."*



(a) Players' Predominant Shot Types (b) Players' Positions

Figure 2.10: Average distance errors for participants' responses to gameplay understanding across two metrics: (a) recognizing players' predominant shot types and (b) identifying players' positions. A Paired t-test revealed that Front Row was significantly ($p < 0.0001$) better than Radio, giving BLV participants an accurate understanding of the gameplay for both metrics. Error bars indicate standard error.

Identifying Players' Positions

Figure 2.10b shows the average error in participants' understanding of players' positions. The mean (± std. dev.) error for Television, Radio, and Front Row was 1.84 (±0.26), 1.49 (±0.49), and 0.41 (±0.43), respectively. The error results for Television failed the Kolmogorov-Smirnov test for normality, i.e., it varied significantly from a normal distribution. Thus, we did not run any parametric tests on the Television error results. A paired t-test was performed on the error results for Radio and Front Row. Average error in participants' responses to players' positions with Front Row was significantly ($t_{11} = 8.04, p < 0.0001$) lower than those with Radio. This suggests that Front Row gave BLV participants a more accurate understanding of the players' positions compared to Radio.

For Television, most participants (n=11) noted that they did not get any useful information about players' positions from the clips, constantly opting for 'I don't know.' P12's response after one of these questions represents participants' overall sentiment: *"Worse than I don't know, no clue"* (P12). For Radio, participants noted that players' positions was rarely specified by the announcers. Even when it was specified, participants expressed difficulties in relating players to their actions:

> P3: *"I believe it wasn't super clear because when [the announcer] said that someone got close to the net, it could have been either one of [the players]."*

With Front Row, participants felt comfortable specifying the players' positions and found that the ability to constantly track players' movements helped them also identify '*when*' a player moved closer to the net:

> P7: *"For the most part, [the players] were far from the net. The left player got close to the net near the end."*

2.6.3 Overall Experience

Here we report our findings for participants' overall experience of viewing tennis games via the three audio format in terms of their perceived information overload, frustration, and immersion. Through these metrics, we aim to quantitatively understand how each audio format fares in terms of our design goal, **G2**: *Providing an appropriate amount of information to facilitate immersion*, which we learned from our formative study (Section 2.2).

Perceived Information Overload

Figure 2.11a shows participants' average TLX scores (1–20, where lower is better) for their perceived information overload for each audio format. The mean (± std. dev.) rating of perceived information overload for Television, Radio, and Front Row were 3.35 (±3.72), 9.78 (±3.19), and 7.38 (±4.64), respectively. A one-way Analysis of Variance (ANOVA) revealed that the audio format has a significant main effect on the perceived information overload ($F_{2,22} = 15.5$, $p < 0.0001$). Post-hoc Turkey test showed that the differences were significant ($p < 0.01$) for every pair of audio format except Radio vs. Front Row.

During the semi-structured interview, we asked participants to elaborate on their information overload scores. Regarding Television, we found that although it was rated to have the least amount of information overload of the three conditions, that fact came at a cost — it did not provide much information at all:

> P7: *"There is no information. Like, you can't overload on what's not there."*

Radio, on the other hand, was noted to provide *"a lot of information to process all at once and really fast"* (P3). P5 further explained:

> P5: *"There's so much being talked about in very little time. And so it doesn't leave a whole lot of room to really ascertain what exactly is happening. It's a lot to take in."*

Front Row was rated by participants as the audio format with the least amount of information overload. However, it was *"kind of overwhelming, at first"* (P3) for participants to get used to Front Row's audio cues. But as participants listened to more clips, Front Row started to feel more intuitive:

> P5: *"Now that I've had three or four different clips [...] it doesn't feel as demanding. And so it's kind of taking on a more natural approach of listening to it."*

Perceived Frustration

Figure 2.11b shows participants' average TLX scores (1–20) for their perceived frustration with each audio format. The mean (± std. dev.) rating of perceived frustration was 15.87 (±5.52) for Television, 9.97 (±5.55) for Radio, and 6.58 (±4.51) for Front Row. The audio formats have a significant main effect on participants' frustration ($F_{2,22} = 11.84$, $p < 0.0003$). Pairwise mean comparison showed the differences were significant between Television and Radio ($p < 0.05$) and between Television and Front Row ($p < 0.01$). However, there was no significant difference between Radio and Front Row in the post-hoc analysis.

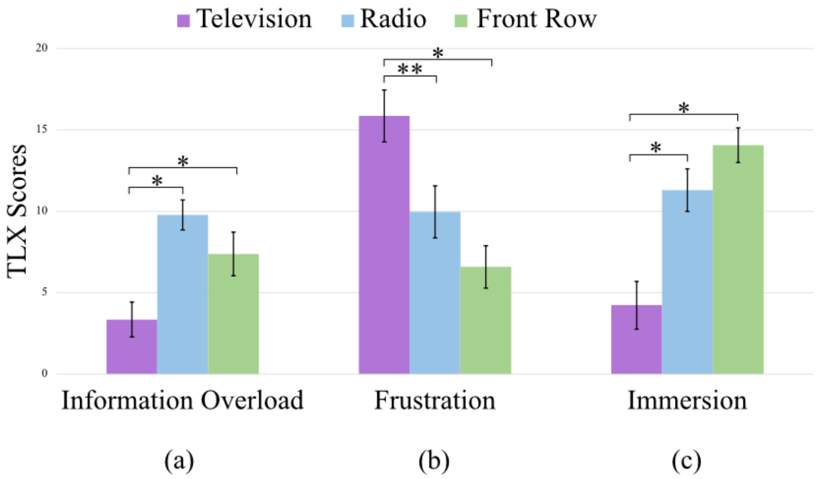


Figure 2.11: Average TLX scores for participants' overall experience. Participants rated their (a) perceived information overload, (b) perceived frustration, and (c) perceived immersion on a scale of 1—20 while viewing tennis rallies via the three audio formats. The error bars indicate standard error. Pairwise significance is depicted for $p < .01$ (∗) and $p < .05$ (∗∗). Participants rated Front Row as the most immersive tennis viewing experience of the three audio formats.

The semi-structured interview allowed us to identify specific aspects of each format that caused the frustration. For Television, most participants (n=11) agreed that their inability to infer gameplay in any meaningful way was frustrating. P11 remarked that she *"couldn't tell what was going on. And when you don't know what's going on, you get frustrated."* P7 also felt strongly about this, exclaiming:

> P7: *"Oh, god, that's a straight 20 [frustration score]. Like, this is the thing that I would change the channel for."*

With Radio, participants expressed frustration about the lack of consistency in the announcer's description of the game. For instance, participants noted that announcers may choose to not describe certain parts of the rallies — for example, completely ignoring one of the players in one rally:

> P12: *"I did feel kinda like I lost out on what was happening with the other player."*

Most participants agreed that this was typical of radio broadcasts, including for sports other than tennis.

Front Row was rated as being the least frustrating of the three audio formats. However, participants found it frustrating to not be able to more accurately discern player movement, specifically along the baseline:

> P11: *"I can tell they're moving [along the baseline], but I just can't get that spatial differentiation on the movement."*

Perceived Immersion

Figure 2.11c shows participants' average TLX scores (1–20) for their perceived immersion while viewing tennis rallies using each audio format. The mean (± std. dev.) rating of perceived immersion were 4.23 (±5.09), 11.30 (±4.53), and 14.07 (±3.70) for Television, Radio, and Front Row, respectively. One-way ANOVA revealed that the audio formats have a significant main effect on participants' immersion within the game ($F_{2,22} = 20.60$, $p < 0.0001$). Post-hoc analysis showed the differences were significant ($p < 0.01$) for all pairs of audio formats except for Radio vs. Front Row.

The semi-structured interview gave us further insight about participants' immersion scores. Most participants stated that Television was not at all engaging.

> P7: *"[Television] is just so under stimulating. It's like if it was between that and a silent room, I would genuinely choose the silent room."*

With Radio, participants felt more immersed because the announcers continuously describe the game, keeping them *"in the game"* (P2). However, the announcers' inability to provide gameplay information in sync with the game, i.e., lagging behind the actual events in the game, derailed participants' sense of immersion:

> P3: *"I don't like that [radio announcers] can't keep up. I don't like that."*

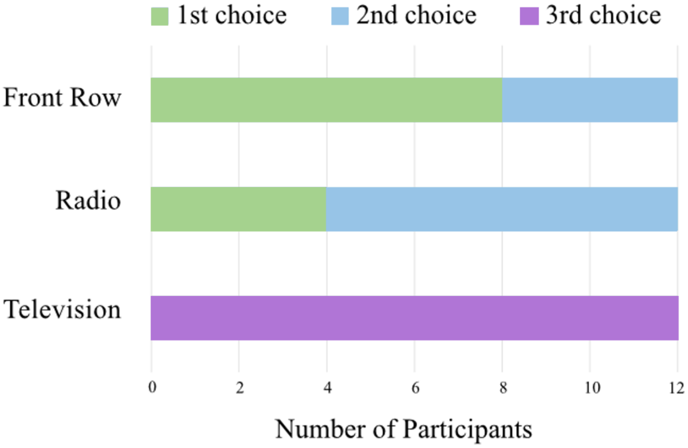


Figure 2.12: Forced ranking results. Participants ranked the three audio formats in order of their preference for viewing tennis games. Eight participants selected Front Row as their number one choice, four participants picked Radio as their top choice, while Television was unanimously ranked as the least preferred option by all participants.

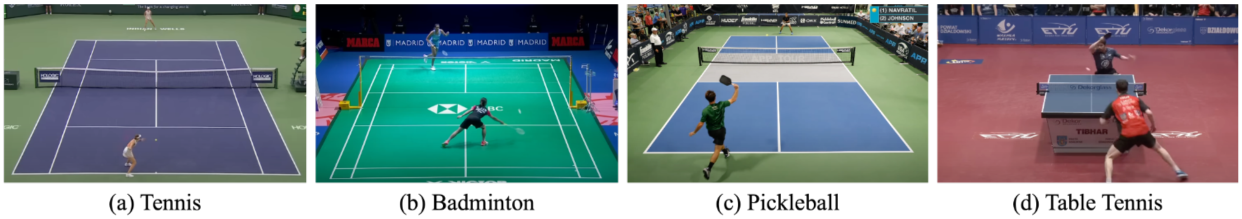

(a) Tennis (b) Badminton (c) Pickleball (d) Table Tennis

Figure 2.13: Illustration of popular racket sports. Front Row's design for (a) tennis can be extended to make other racket sports, such as (b) badminton, (c) pickleball, and (d) table tennis, accessible to BLV viewers.

Front Row was rated as the most immersive of the three audio formats, with most participants appreciating how it renders the gameplay in a spatial manner:

> P3: *"[What] I liked about [Front Row] was being able to hear objects in space, which is really important. So, you know the players and their movement. I think it is really fascinating. And that's something that is oftentimes missed."*

Most Participants were excited about how Front Row made them feel more *"involved in what was going on"* (P11) in the rally, by giving them the ability to follow the game in sync with the actual events:

> P11: *"[In Radio], the announcer lags behind. So, [with Front Row] I like the real-time [aspect] of knowing the types of shots that are shot at the time."*

However, the use of synthetic audio cues in Front Row negatively affected some participants' (n=3) sense of immersion.

2.6.4 Forced Ranking Results

Figure 2.12 shows how participants ranked the audio formats in order of their preference for watching tennis games. Eight out of twelve participants chose Front Row as their preferred audio format and four chose Radio. All participants unanimously rated Television as their least preferred format. Half of the participants who rated Radio as their number one choice (n=2) acknowledged that Front Row was a "*close second*" (P12).

In the semi-structured interview, we asked participants to elaborate on their rankings. For Radio, it was the announcers' ability to convey the emotions of the game that caused participants to rate radio favorably. For example, participants liked how announcers "*put a flair on everything so that it could sound interesting. [Announcers] put character into each player.*" For Immersive, it was the ability gain a spatial understanding of the gameplay in an immersive manner, as well as its ability to offer them agency in interpreting the gameplay for themselves and forming their own opinions about the players' moods and strategies during the game:

> P5: *"As a blind person, oftentimes, descriptions are through the lens of how other people perceive things. Having that information conveyed just in its most raw and basic form allows me to [...] make connections that I can derive on my own."*

2.7 Discussion

Our goal with Front Row was to explore the idea of making sports broadcasts accessible by generating immersive audio representations directly from the source videos. Similar to previous work in sports video analysis (Section 2.1.3), our approach uses computer vision to give our system an understanding of what is happening in the game. Unlike previous work, however, our approach focuses on sharing that understanding with *people* who could benefit from it via immersive audio cues that we designed. We reflect upon the implications of this approach for ongoing work in visual media accessibility (Section 2.1.1), sports broadcast accessibility (Section 2.1.2), and sports video analysis via computer vision (Section 2.1.3).

Implications for visual media accessibility

Regarding the more general problem of visual media accessibility, our approach represents a shift away from textual descriptions and toward a more direct representation of raw visual details—for example, continuously displaying players' raw positions rather than describing players' positions via speech. Our results show this shift has many advantages for BLV people, including giving them a better understanding of players' positions (Figure 2.10) and making that understanding real-time rather than time-delayed as with radio announcer's descriptions. Our results also show, however, that this shift is not a complete replacement for textual descriptions. One-third of participants preferred radio over Front Row (Figure 2.12), and there was no significant difference in feeling of immersion between radio and Front Row (Figure 2.11).

A major reason for this is that textual descriptions can convey important story elements or contextual details that lie beyond what we captured via our computer vision-based approach. Radio announcers might mention, for example, that a player's forehand has been really strong all year and that it is great that the player has been playing a lot of forehands. Computer vision alone cannot capture this type of broader context.

As a result, we have found a need for more immersive approaches to visual media accessibility as well as a need for understanding textual descriptions' unique affordances. Future work in visual media accessibility (including images as well as more work on video) can explore designs that combine textual descriptions with immersive representations to realize the advantages of both. In this process, sound design will be very important. From our design process and studies, we learned that users feel less immersed when an immersive representation's sound cues seem artificial. Front Row's representation of player's positions would have been stronger, for example, if it used footstep sounds rather than continuous humming sounds.

Front Row presents an approach for designing spatialized (3D) audio cues to make videos — specifically tennis videos— more accessible and immersive to BLV people. Future research could explore how spatialized audio cues should be used for making other, more interactive, types of visual media such as video games and virtual reality (VR) applications accessible to BLV people.

Implications for sports broadcast accessibility

Front Row demonstrates that it is possible to make tennis broadcasts accessible to BLV people without extensive hardware installations, as required by prior work such as Action Audio [21]. A direct implication of this is that other racket sports such as badminton, pickleball, table tennis, and squash (collectively shown in Figure 2.13) could be made accessible to BLV people by training sport-specific computer vision pipelines and using Front Row's overall approach.

Our work leaves open questions, however, about how to make sports with larger fields and more players (sports such as football and basketball) accessible. TV broadcasts for tennis often employ a single, fixed camera position that covers the entire court. For such other sports, however, that is not the case—the camera cuts between many different views, and the court or field is very rarely shown on screen in its entirety. For a Front Row-like approach to be effective with these other sports, its computer vision pipeline would need to be evolve to fuse many camera views into a single, cohesive field representation. Another challenge will be to convey the positions and actions of many players on the field without overloading the viewer. Prior work on tactile graphics for football [74] can inspire future work on addressing this challenge.

Last, we learned that BLV people greatly value watching sports with friends and family. Most work in sports broadcast accessibility, however, has focused on developing novel user interfaces and evaluating them in the context of BLV users watching sports by themselves. Future research in this space should also evaluate their ability to help BLV people in social scenarios such as watching with friends and family on a couch at home.

2.8 Future Directions for Immersive Sports Audio

Front Row's current design introduces a number of opportunities for future work:

Conveying intensity of play via multimodal representations

Front Row supports BLV viewers' understanding of gameplay via audio representations of players' positions and types of shots. These representations currently do not provide viewers with information on variations in players' running speeds and intensity of shot-making. Access to the subtle intensity variations in the game could help BLV viewers gain insights into more abstract aspects of gameplay, such as players' characters and emotions, and further enhance their ability to visualize the action. In the future, we will investigate how to convey the intensity of play to BLV viewers in a manner that does not increase their information overload. One possible solution to convey more information without overloading users is to use multimodal representations. For example, a smartwatch worn on the wrist could convey how hard players hit the ball via haptic feedback, in sync with Front Row's bell sound cues for shots.

Supporting different viewer expertise levels

Front Row currently uses a fixed set of audio representations for conveying the players' positions and shots. However, participants indicated different preferences based on their familiarity with tennis. For instance, participants who self-reported as "*expert viewers*" (rated 4+ on tennis familiarity) craved more fine-grained and technical information on players' shots, such as learning whether the forehand was a top-spin, a volley, or a lob. Expert viewers also wanted more control over the parameters of the sound cues, such as pitch and volume. “Casual viewers” (rated 1–3 on tennis familiarity), by contrast, preferred Front Row’s current configuration with only two shot types because “*putting too many sounds would become very confusing*” (P11) for them. In the future, we will investigate these differing preferences between viewers of different expertise levels in order to allow BLV viewers to customize their 3D soundscapes in Front Row. We envision introducing different modes in Front Row that viewers could select based on their expertise, with the ability to fine-tune these baseline configurations as per individual preferences.

Implications for BLV athletes in coaching and strategy

We designed Front Row for BLV sports viewers, but future work could investigate how Front Row can be developed further to support blind athletes in coaching and learning strategies. Professional athletes review video footage of themselves in order to improve their techniques and also of their opponents to identify opponents’ playing styles, weaknesses, and strengths [96]. To support these affordances for BLV athletes, future research could investigate what information athletes want from their video analysis and design audio representations that effectively render this information to them. BLV participants who play blind tennis [28] (P4, P5, P8) expressed excitement about sharing their experience of using Front Row with their coaches and friends for this purpose.

2.9 Applications

We now illustrate applications that Front Row could enable in the future. Figure 2.14 shows a Front Row plug-in for video streaming platforms to make sports videos across the Web accessible. As Figure 2.15 shows, Front Row can make recreational tennis games at high schools, parks, and universities accessible to BLV audiences by processing video feed from a camera on court. Figure 2.16 illustrates Front Row’s potential in making video games accessible.

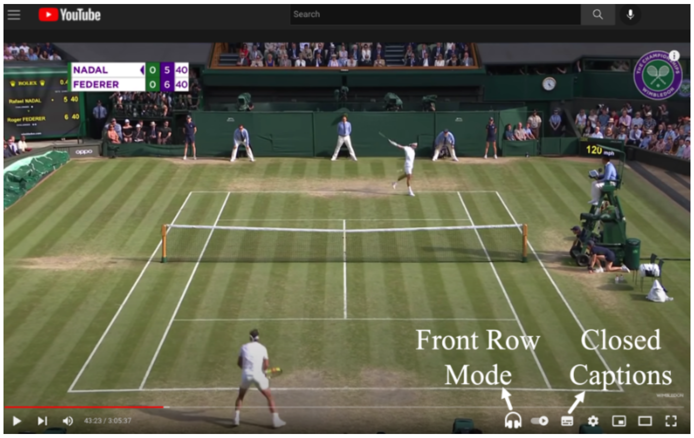


Figure 2.14: Front Row plug-in for video streaming platforms. Front Row can be integrated with online video streaming platforms, such as YouTube, ESPN+, and Hulu, to make recorded tennis broadcasts accessible to BLV viewers. This could work similarly to how closed captions are implemented on YouTube. Video source: Wimbledon's YouTube channel.

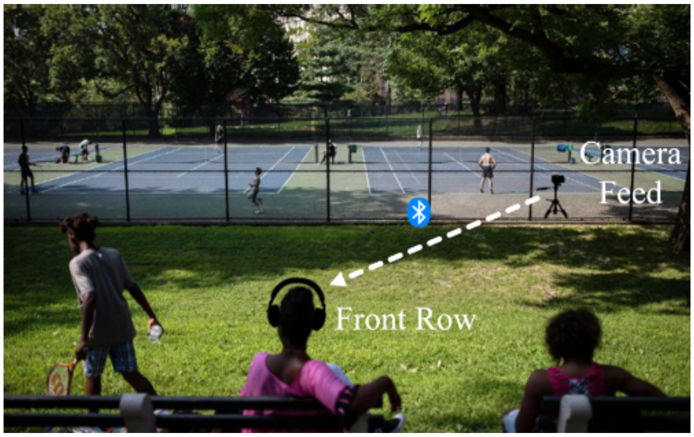


Figure 2.15: Recreational tennis game at a park. Front Row can make recreational tennis matches, such as matches at high schools, parks, and universities, accessible to BLV viewers. By processing a camera feed captured behind one of the players, Front Row can enable BLV audience members to follow the game in real time. Image source: The New York Times.

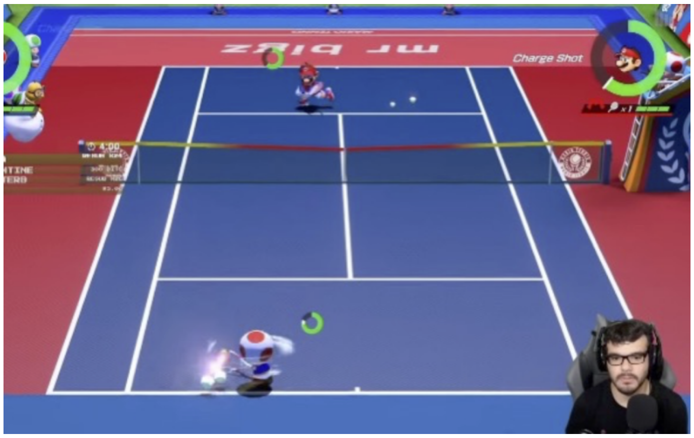


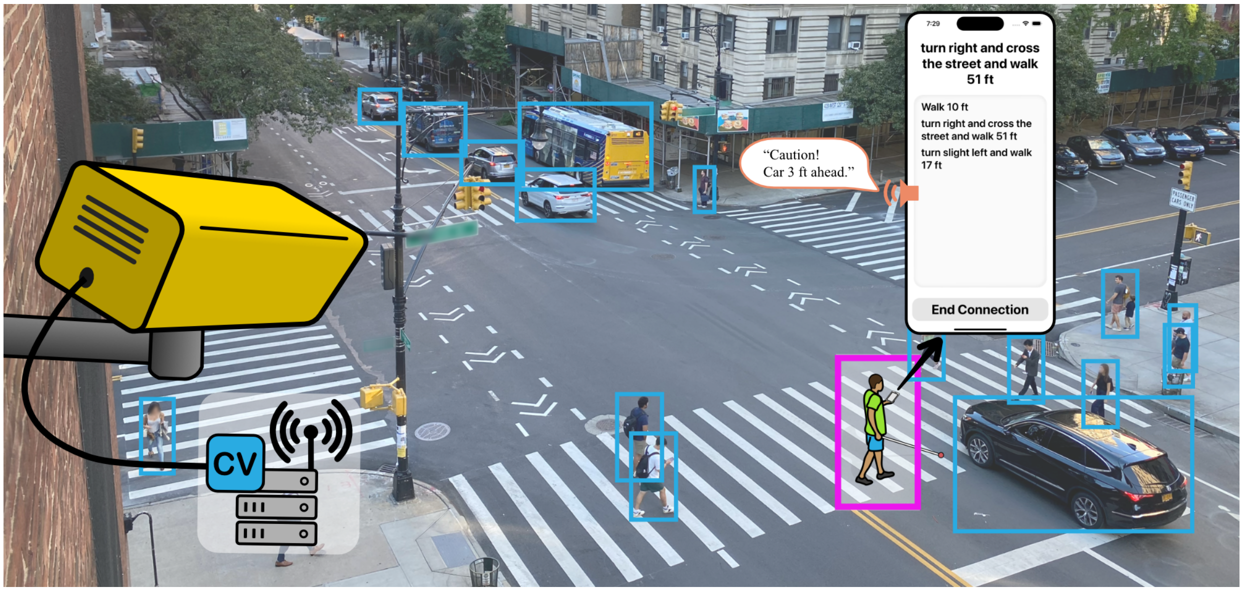
Figure 2.16: Gaming streams. A tennis video game stream by *Ray*, a popular streamer, who can be seen on the bottom right playing *Mario Tennis Aces*. Front Row can make both video game streams and video games themselves accessible to BLV viewers and gamers. Video source: Ray's YouTube channel.

Chapter 3: StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians

Outdoor navigation in unfamiliar environments is a major challenge for blind and low-vision (BLV) people. Among the many navigation systems that have been developed to assist BLV people outdoors, GPS-based systems are the most popular [97, 98, 99, 100, 101]. These systems, such as BlindSquare [97] and Microsoft Soundscape [98], guide users to a destination and notify them of surrounding points of interest (POIs). Despite GPS's undeniable impact in making outdoor environments navigable, its imprecision is a major limitation [102]. GPS precision can range from 5 meters at best to over tens of meters in urban areas with buildings and trees [103, 104, 105]. This imprecision causes BLV people to veer off track [106], run into unexpected obstacles [107, 108, 109], and struggle to reach precise destinations [102] when navigating outdoors.

Prior work on indoor navigation, on the contrary, has made precise navigation assistance possible for BLV people [110, 111, 112, 113, 114]. Most approaches do so by installing a dense network of additional hardware, such as Bluetooth [110] or WiFi [112] beacons, to precisely locate a user's position. Retrofitting outdoor environments with additional hardware, however, is not feasible due to the vast scale and complex nature of outdoor spaces. It would require extensive financial investments and coordination with city authorities to install and maintain such specialized hardware, which may not be possible.

Ironically, many outdoor environments of interest, such as urban districts and downtown areas, are already instrumented with hardware that has the potential to help, including street cameras, traffic sensors, and other urban infrastructure components. Street cameras, in particular, are increasingly being installed in cities for public safety, surveillance, and traffic management-related applications [115, 116, 117, 118, 119]. Although these pre-existing street cameras have been de-



turn right and cross the street and walk 51 ft

Walk 10 ft

turn right and cross the street and walk 51 ft

turn slight left and walk 17 ft

Figure 3.1: StreetNav is a system that explores the concept of repurposing existing street cameras to support precise outdoor navigation for blind and low-vision (BLV) pedestrians. It comprises two components: (i) a computer vision (CV) pipeline, and (ii) a companion smartphone app. The computer vision pipeline processes the street camera's video feeds and delivers real-time navigation feedback via the app. StreetNav offers precise turn-by-turn directions to destinations while also providing real-time, scene-aware assistance to prevent users from veering off course, alert them of nearby obstacles, and facilitate safe street crossings.

ployed for purposes unrelated to accessibility, their potential for facilitating navigation assistance for BLV people remains largely untapped.

In this work, we explore the idea of leveraging existing street cameras to support outdoor navigation assistance, and we investigate the effectiveness of such an approach.

We seek to answer the following research questions:

RQ1. What challenges do BLV people face when navigating outdoors using GPS-based systems?

RQ2. How should street camera-based systems be designed to address BLV people's challenges in outdoor navigation?

RQ3. To what extent do street camera-based systems address BLV people's challenges in outdoor navigation?

To answer RQ1, we conducted formative interviews with six BLV pedestrians and discovered the challenges BLV people face when navigating outdoors using GPS-based systems. Our participants reported challenges in following GPS's routing instructions through complex environment layouts, avoiding unexpected obstacles while simultaneously using assistive technology, and crossing streets safely.

To answer RQ2, we developed *StreetNav*, a system that leverages a street camera to support precise outdoor navigation for BLV pedestrians. As Figure 3.1 illustrates, StreetNav comprises two key components: (i) *a computer vision pipeline* and (ii) *a companion smartphone app*. The computer vision pipeline processes the street camera's video feed and delivers real-time navigation assistance to BLV pedestrians via the smartphone app. StreetNav offers precise turn-by-turn directions to destinations while also providing real-time, scene-aware assistance to prevent users from veering off course, alert them of nearby obstacles, and facilitate safe street crossings. We developed StreetNav at the NSF PAWR COSMOS wireless edge-cloud testbed in New York City.

StreetNav uses one of the testbed's street cameras mounted on the second floor of a building, which faces a four-way street intersection.

To answer RQ3, we conducted user evaluations involving eight BLV pedestrians who navigated routes with both StreetNav and BlindSquare [97], a popular GPS-based navigation app especially designed for BLV people. Our findings reveal that StreetNav offers significantly greater precision in guiding pedestrians compared to BlindSquare. Specifically, StreetNav guided participants to within an average of 2.9 times closer to their destination and reduced veering off course by over 53% when compared to BlindSquare. This substantial improvement was reflected in the unanimous preference of all participants for StreetNav over BlindSquare in a forced ranking. Our evaluation, however, also revealed technical considerations related to StreetNav's performance, notably its sensitivity to lighting conditions and environmental occlusions. We discuss the future implications of our findings in the context of deploying street camera-based systems at scale for outdoor navigation assistance.

In summary, we contribute (1) a formative study of BLV people's challenges in outdoor navigation using GPS-based systems, (2) the StreetNav system through which we explore the concept of repurposing street cameras for precise outdoor navigation assistance, and (3) a user evaluation of StreetNav.

3.1 Related Work

Our work builds on the following three main research threads: (i) outdoor navigation approaches, (ii) overhead camera-based robot navigation, and (iii) indoor navigation approaches.

Outdoor Navigation Approaches

Existing approaches for outdoor navigation primarily rely on GPS-based navigation systems for guiding users to the destination and providing information about nearby POIs [97, 98, 99, 100, 101]. BlindSquare[97], for instance, utilizes the smartphone's GPS signal to determine the user's location and then provides the direction and distance to the destination, gathered from Foursquare and Open Street Map. The GPS signal, however, offers poor precision with localization errors as big as tens of meters [104, 120, 110, 103]. The accuracy is lower in densely populated cities [121], which is even more concerning given that a disproportionately high percentage of BLV people live in cities [122]. Despite GPS-based systems' undeniable impact on helping BLV people in outdoor navigation, their low precision and inability to provide real-time support for avoiding obstacles and veering off the path limits their usability as a standalone navigation solution. Our work attempts to investigate street cameras' potential as an alternative solution for providing precise and real-time navigation assistance.

Another approach for outdoor navigation has explored developing personalized, purpose-built, assistive devices that support BLV people with scene-aware aspects of outdoor navigation, such as crossing streets [123, 124, 125], recording routes [120], and avoiding obstacles [126, 127, 128, 129, 130, 131]. While these solutions address some of the precise and real-time aspects of BLV people's outdoor navigation, support for point-to-point navigation is missing. Consequently, they do not offer a comprehensive, all-in-one solution for outdoor navigation. Furthermore, these systems place the burden of purchasing costly devices onto the BLV users. Our work, by contrast, explores the possibility of using existing street cameras to provide a comprehensive solution for outdoor navigation. We investigate repurposing existing hardware in outdoor environments to support accessibility applications, thus imbuing accessibility within the city infrastructure directly, and adding no additional cost to the BLV user.

Overhead Camera-based Robot Navigation

A parallel research space to street cameras for blind navigation is robot navigation using overhead cameras. One common subspace within this field is sensor fusion for improved mapping. Research in this space focuses on fusing information between sighted "guide" robots and overhead cameras [132], fusing multiple camera views for improved tracking [132, 133, 134], and improving homography for robust mapping, independent of camera viewing angle [135, 136]. Another challenge tackled within this space is robot path planning. Research in this space aims to improve path planning algorithms [132, 133, 136], assign navigational tasks to robot assistants [132, 133], and address the balance between obstacle avoidance and path following [132, 136]. While prior work on robot navigation using fixed cameras explores the research space of automating “blind” robot navigation, our work explores how fixed cameras, specifically street cameras, could be repurposed to support navigation for blind pedestrians. Our work considers BLV users’ needs and preferences around outdoor navigation to design and develop a system that can offer precise navigation assistance.

Indoor Navigation Approaches

Prior work in indoor navigation assistance has made significant progress through the utilization of various localization technologies, which usually relies on retrofitting the environment with additional hardware like WiFi or Bluetooth beacons [110, 111, 112, 113, 114]. These solutions have proven highly effective within indoor environments. NavCog3 [110], for example, excels in indoor navigation by employing Bluetooth beacons for precise turn-by-turn guidance. Nakajima and Haruyama [111] exploit the use of visible lights communication technology, utilizing LED lights and a geomagnetic correction method to localize BLV users. However, extending these approaches to support outdoor navigation is not practical. This is particularly evident when considering the substantial initial investment in hardware setup that these systems typically require, making them ill-suited for the larger, unstructured outdoor environment. Furthermore, most of these methods lack the capability to assist with obstacle avoidance and to prevent users from veering off course — both of which are less severe issues indoors compared to outdoors [106]. In contrast, our exploration of using existing street cameras is better suited to address the largely unaddressed challenge of outdoor pedestrian navigation. This approach offers precise localization without requiring supplementary hardware, harnessing street cameras for locating a pedestrian’s position. Additionally, it holds the potential to effectively tackle the distinctive challenges posed by the unstructured nature of outdoor environments, including real-time obstacle detection and the interpretation of critical visual cues like street crossing signals.

3.2 Formative Interviews

We conducted semi-structured interviews with six BLV participants to identify BLV pedestrians' challenges in outdoor navigation when using GPS-based systems (RQ1).

3.2.1 Methods

Participants

We recruited six BLV participants (three males and three females, aged 29–66) by posting on social media platforms and snowball sampling [94]. Table 3.1 summarises the participants' information. All interviews were conducted over Zoom and lasted about 90 minutes. Participants were compensated $25 for this IRB-approved study.

Procedure

To identify the specific challenges that BLV people face when navigating outdoors, we used a recent critical incident technique (CIT) [81], in which we asked participants to recall and describe a recent time when they navigated outdoor environments using GPS-based assistive technology (AT). For example, we first asked participants to name the AT they commonly use and then asked them to elaborate on their recent experience of using it: *"So, you mentioned using BlindSquare a lot. When was the last time you used it?"* Then, we initiated a discussion by establishing the scenario for them: *"Now, let's walk through your visit from the office to this restaurant. Suppose, I spotted you at your office. What would I observe? Let's start with you getting out of your office building."* We asked follow-up questions to gain insights into what made the aspects of outdoor navigation challenging and what additional information could help address them.

Interview Analysis

To analyze the interviews, we first transcribed the study sessions in full and then performed thematic analysis [82] involving three members of our research team. Each researcher first inde-

Table 3.1: Self-reported demographics of our participants. Gender information was collected as a free response; our participants identified themselves as female (F) or male (M). Participants rated their assistive technology (AT) familiarity on a scale of 1–5.

PID	Age	Gender	Race	Occupation	Vision ability	Onset	Mobility aid	AT familiarity (1–5)
F1	29	Female	White	Claims expert	Totally blind	At birth	White cane	3: Moderately familiar
F2	61	Female	White	Retired	Light perception only	Age 6	Guide dog	1: Not at all familiar
F3	66	Female	White	Retired	Totally blind	Age 58	Guide dog	2: Slightly familiar
F4	48	Male	Black	Unemployed	Light perception only	Age 32	White cane	3: Moderately familiar
F5	27	Male	Mixed	Unemployed	Totally blind	At birth	White cane	3: Moderately familiar
F6	38	Male	White	AT instructor	Totally blind	At birth	White cane	5: Extremely familiar

pendently went through the interview transcripts and used NVivo [83] to create an initial set of codes. Then, all three iterated on the codes together to identify emerging themes.

3.2.2 Findings: BLV Pedestrians' Challenges in Outdoor Navigation

We found three major themes around challenges that BLV pedestrians face when navigating outdoors using GPS-based systems.

C1: Routing through complex environment layouts

GPS-based systems, such as BlindSquare [97], offer navigation instructions that follow a direct path to the destination from the user's current position, often referred to as "as the crow flies," rather than providing detailed turn-by-turn instructions through a poly-line path that guide BLV people through the environment layout. Since *"not everything is organized in the ideal grid-like way"* (F1), participants reported difficulties following the "as the crow flies" instructions, failing to confidently act upon the instructions without any knowledge of the environment layout. This was particularly challenging in complex layouts, as F3 recalled: *"I didn't know if crosswalks were straight or curved or if they were angled. [It was hard] to figure out which way you needed to be to be in the crosswalk."* Many participants cited problems such as making the wrong turns into unexpected "alleyways" (F1, F2, F4) that landed them in dangerous situations with "cars coming through" (F2). Participants cited examples about how these instructions were often inaccurate, causing them to veer off course—a common issue for BLV people in open, outdoor space [106]—

and end up in the middle of the streets.

C2: Avoiding unexpected obstacles while using GPS-based systems

BLV people's challenges relating to obstacles during navigation are well researched [107, 108]. However, we found specific nuances in their difficulties, particularly when they rely on their conventional mobility aids in conjunction with GPS-based navigation systems. Participants commonly reported the use of mobility aids like white canes alongside GPS systems for guidance. During this combined navigation process, they encountered difficulties in maintaining their focus on obstacle detection, often resulting in collisions with objects that they would have otherwise detected using their white canes. For instance, F2 shared an incident where they remarked, *"there were traffic cones [and] I tripped over those"* while following directions. Notably, moving obstacles such as pedestrians and cars, as well as temporarily positioned stationary obstacles like triangle sandwich board signs, posed significant challenges for navigation. F4 expressed this sentiment, stating, *"You know how many times I've walked into the sides of cars even though I have the right of way. Drivers have gotten angry, accusing me of scratching their vehicles. It can spoil your day [and make] you feel insecure."*

C3: Crossing street intersections safely

Consistent with prior research [125, 137, 138], our study participants highlighted that crossing streets remained a significant challenge for them. Since GPS-based systems do not help with street-crossing, most participants relied on their auditory senses. They mentioned the practice of listening for vehicular sounds to gauge traffic flow on streets running parallel and perpendicular to their position. This auditory technique helped them assess when it was safe to cross streets. However, participants also reported instances where this method proved inadequate due to external factors: *"yeah, it can be tricky, because [there may be] really loud construction nearby that can definitely throw me off because I'm trying to listen to the traffic"* (F1). Furthermore, their confidence in street-crossing decisions was affected by their inability to ascertain the duration of pedestrian signals and the length of the crosswalk. This uncertainty led to apprehension, as they expressed a fear of becoming stranded mid-crossing, as exemplified by one participant's comment: "*I don't want to be caught in the middle [of the street]*" (F4).

3.3 The StreetNav System

StreetNav is a system that explores the concept of repurposing street cameras to support outdoor navigation for BLV pedestrians (RQ2, RQ3). It provides users precise turn-by-turn navigation instructions to destinations (**C1**), helps prevent veering off track (**C1**), gain awareness of nearby obstacles (**C2**), and assist in crossing streets safely (**C3**). StreetNav enables these navigation affordances through its two main components: (i) *computer vision pipeline*, and (ii) *companion smartphone app*. The computer vision pipeline processes the street camera's video feeds to give BLV pedestrians real-time navigation feedback via the app. Our design and development of StreetNav considers prior work on navigation assistance, functions of traditional mobility aids, and formative interviews with BLV people (Section 3.2) that identified challenges they face when navigating outdoors using existing GPS-based systems.

The following sections describe StreetNav's technical setup (Section 3.3.1), the computer vision pipeline (Section 3.3.2), and the smartphone app's user interface (Section 3.3.3).

3.3.1 StreetNav: Technical Setup

Figure 3.2 shows the street camera we used for developing and evaluating StreetNav. We chose this camera because it faces a four-way street intersection—the most common type of intersection—and is mounted on a building's second floor, offering a typical street-level view of the intersection. The camera is part of the NSF PAWR COSMOS wireless edge-cloud testbed established by our research team.

StreetNav's computer vision pipeline takes the real-time video feed from the camera as input. For this purpose, we deployed the computer vision pipeline on one of the testbed servers, which captures the camera's video feed in real time.

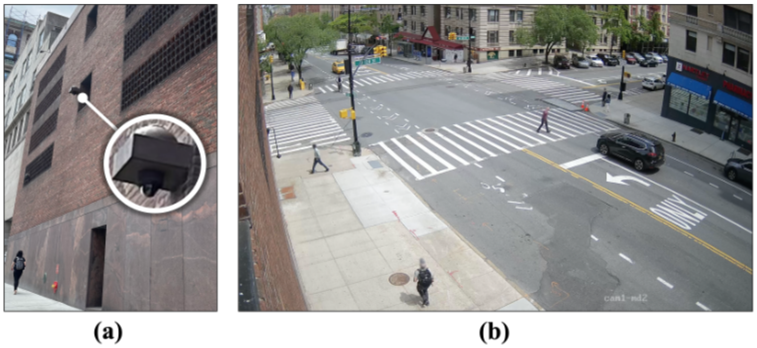


Figure 3.2: Street camera used for StreetNav's development and evaluation. The camera (a) mounts on the building's second floor, and (b) captures the view of a four-way intersection.

This server runs Ubuntu 20.04 with an Intel Xeon CPU@2.60GHz and an Nvidia V100 GPU.

StreetNav's two components—the computer vision pipeline and the app—interact with each other via a cloud server, sharing information using the MQTT messaging protocol [139]. Since MQTT is a lightweight messaging protocol, it runs efficiently even in low-bandwidth environments. The computer vision pipeline only sends processed navigation information (e.g., routing instructions, obstacle's category and location) to the app, rather than sending video data. This alleviates any privacy concerns around streaming the video feed to the users and avoids any computational bottlenecks that may happen due to smartphones' limited processing capabilities. It is worthwhile to note that although we have fulfilled the necessary anonymization and privacy measures for this prototype version of StreetNav, the deployment of such a system at scale would require a deeper exploration into privacy and ethics.

The StreetNav app's primary purpose is to act as an interface between the user and the computer vision pipeline. We developed StreetNav's iOS App using Swift [140], enabling us to leverage VoiceOver [141] and other built-in accessibility features.

3.3.2 StreetNav: Computer Vision Pipeline

StreetNav's computer vision pipeline processes the street camera's video feed in real time to facilitate navigation assistance. It consists of four components: (i) *localizing and tracking the user*: locating user's position on the environment's map; (ii) *planning routes*: generating turn-by-turn navigation instructions from user's current position to destinations; (iii) *identifying obstacles*: predicting potential collisions with other pedestrians, vehicles, and objects (e.g., trash can, pole); and (iv) *recognizing pedestrian signals*: determining when it is safe for pedestrians to cross (walk vs. wait) and calculating the duration of each cycle. Next, we describe the computer vision pipeline's four components in detail.

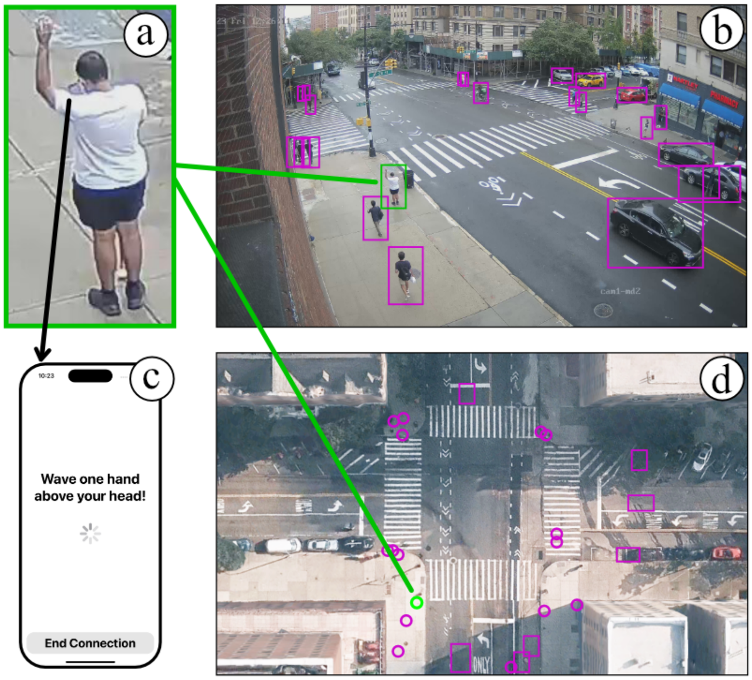
Localizing and tracking the user

To offer precise navigation assistance, the system must first determine the user's position from the camera view and then project it onto the environment's map. Figure 3.3d shows the map representation we used, which is a snapshot from Apple Maps' [142] satellite view of the intersection where the camera is deployed.

StreetNav tracks pedestrians from the camera's video feed using Nvidia's DCF-based multi-object tracker [143] and the YOLOv8 object detector [144]. The computer vision pipeline is developed using Nvidia GStreamer plugins [145, 146], enabling hardware-accelerated video processing to achieve real-time tracking. We chose this tracker for its trade-off between real-time performance and robustness to occlusions. The tracker detects all pedestrians and assigns them a unique ID. However, the system needs a way to differentiate between the BLV user and other pedestrians.

Figure 3.3 shows the *gesture-based localization* approach we introduced to address this issue. To connect with the system, BLV pedestrians must wave one hand above their head for 2–3 seconds (Figure 3.3a), enabling the system to determine the BLV pedestrian's unique tracker ID. We chose this gesture after discussions with several BLV individuals, including our BLV co-author, and most agreed that this single-handed action was both convenient and socially acceptable to them. Moreover, over-the-head gestures such as waving a hand can also be detected when users are not directly facing the street camera.

StreetNav implements the gesture-based localization approach by first creating image crops of all detected pedestrians and then classifying them as 'waving' or 'walking' pedestrians using CLIP [10]. CLIP classifies each pedestrian by computing visual similarity between the pedestrian's



Wave one hand above your head!

Figure 3.3: Gesture-based localization for determining a user's position on the map. (a) A study participant (P1) is (c) prompted to wave one hand above their head, enabling the computer vision pipeline to distinguish them from other pedestrians in (b) the camera feed view and (d) the map.

image crop and two language prompts: 'person walking' and 'person waving hand.' We experimentally fine-tuned the confidence thresholds and these language prompts. We also tried other action recognition models, such as MMaction2 [147], but found that our CLIP-based approach was much faster and robust to false positives.

Finally, we transformed the user's position on the street camera view (Figure 3.3b) onto the map (Figure 3.3d) using a simple feed-forward neural network, trained on data that we manually annotated. The network takes as input the 2D pixel coordinate from the street camera view and outputs the corresponding 2D coordinate on the map. StreetNav continuously tracks the user from the camera feed and transforms its position onto the map.

Planning routes

StreetNav represents routes as a sequence of straight lines on the map connected by waypoints. To plan routes, StreetNav requires that a map of the environment is annotated with waypoints

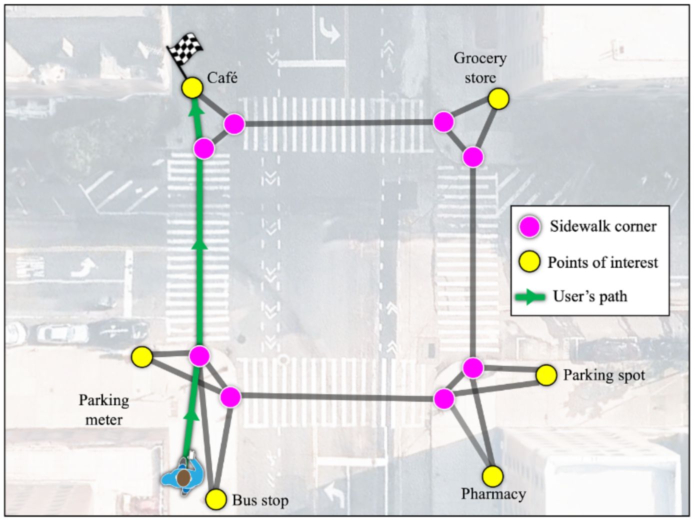


Figure 3.4: StreetNav's internal graph representation for route planning. The user's current position is added dynamically as a start node to the graph upon choosing a destination. The shortest path, highlighted in green, is then calculated as per this graph representation.

and connections between them. This offline process is performed by manually annotating the environment's map, as shown in Figure 3.4. The administrator marks two types of points on the map: POIs and sidewalk corners. The POIs are potential destinations that users can choose from. The sidewalk corners act as intermediary waypoints en route to the destination. We chose sidewalk corners as waypoints because BLV pedestrians often look for the tactile engravings at sidewalk corners to help orient themselves and transition into crosswalks. Thus, these waypoints blend in well with BLV users' current navigation practices.

Figure 3.4 shows the internal graph structure that StreetNav uses for planning routes. This graph-based representation of the environment has also been used in prior work on indoor navigation systems [110, 114, 148]. In the graph, nodes correspond to POIs and sidewalk corners, whereas edges correspond to walkable paths. Once the user chooses a destination from the POIs, StreetNav adds the user's current position as a start node to this graph representation and computes the shortest path to the chosen POI using A* algorithm [149]. Figure 3.4 highlights the shortest path from the user's current position to the chosen destination (café). This route enables StreetNav to guide users to the destination via turn-by-turn instructions.

Identifying obstacles

Prior work on obstacle avoidance developed systems that guide BLV people around obstacles [148, 150]. StreetNav, however, aims to augment BLV pedestrians' awareness of obstacles to help them confidently avoid obstacles using their traditional mobility aids (e.g., white cane) and mobility skills. From our formative interviews, we learned that obstacles that catch BLV users unexpectedly were specifically hard to avoid in outdoor environments (**C2**). Thus, StreetNav provides users with information about the obstacle's category and relative location. This gives BLV users context on the size, shape, and location of an obstacle, enabling them to confidently use their mobility skills around unexpected obstacles.

Figure 3.5 illustrates how the system identifies obstacles in the user's vicinity. StreetNav's multi-object tracker is used to track other objects and pedestrians. Examples of other objects include cars, bicycles, poles, and trash cans. The computer vision pipeline then projects the detected objects' positions onto the map. To identify obstacles in the BLV user's vicinity, StreetNav computes the distance and angle between the user and other detected objects with respect to the map (Figure 3.5b). Any object (or pedestrian) within a fixed radial distance from the BLV user is flagged as an obstacle. Through a series of experiments with our BLV co-author, we found that a 4 foot radius works best for StreetNav to provide users with awareness of obstacles in a timely manner.

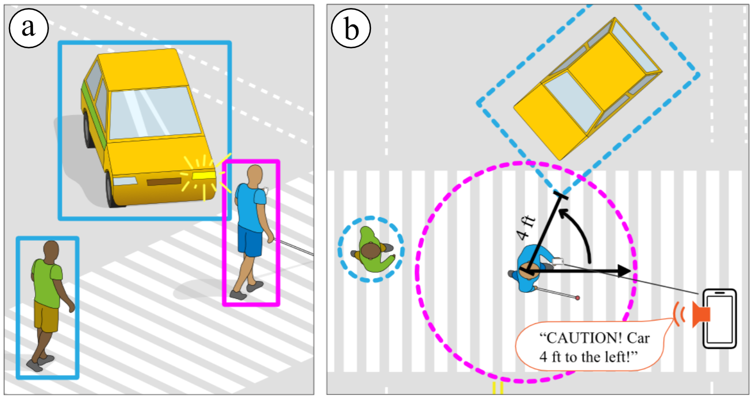


Figure 3.5: Identifying obstacles in the user's vicinity. (a) A vehicle turning left yields to the BLV pedestrian (detected in purple) crossing the street. (b) StreetNav identifies the obstacles' category and relative location on the map to provide real-time feedback via the app.

Recognizing pedestrian signals

To determine the pedestrian signals' state (i.e., *walk* vs. *wait*), we leverage the fact that walk signals are always white, whereas wait signals are always red in color. StreetNav requires the pixel locations of the pedestrian signals in the video feed in order to recognize the signal state. The administrator annotates the video feed image to draw a bounding box around the pedestrian signals' screen. Since the position of pedestrian signals is fixed with respect to the mounted street camera, this process needs to be done only once during setup, along with the map annotation process described earlier.

Figure 3.6 shows the annotated pedestrian signals in the camera's video feed. StreetNav uses these annotations first to generate image crops of the two signals and then threshold both image crops to filter all red and white pixels. It compares the number of white and red pixels in each crop to identify the signal's state: *walk* (Figure 3.6a) vs. *wait* (Figure 3.6b). We experimentally fine-tuned the count thresholds to accurately identify the signal state. Although the two crops are low resolution, this approach still yields accurate results since it distinguishes the state using pixel colors.

Our formative interviews found that BLV pedestrians faced difficulty pacing themselves while crossing streets (**C3**). To address this challenge, StreetNav provides users with information about how much time remains for them to cross. StreetNav's computer vision pipeline computes the time remaining to cross by keeping track of the signal cycles' duration. StreetNav maintains a timer that records the moments when each signal changes its state. After observing a full cycle, StreetNav is able to accurately keep track of both the state and timing of each signal. StreetNav periodically refreshes the timer to adapt to any changes in signal duration that may happen for traffic management reasons.

3.3.3 StreetNav App: User Interface

The StreetNav iOS app interacts with the computer vision pipeline to allow BLV pedestrians to choose a destination and receive real-time navigation feedback that guides them to it. BLV users

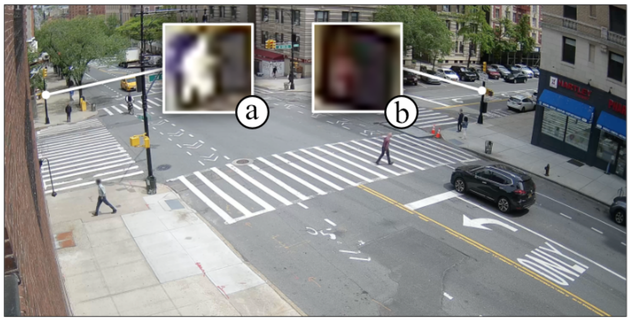


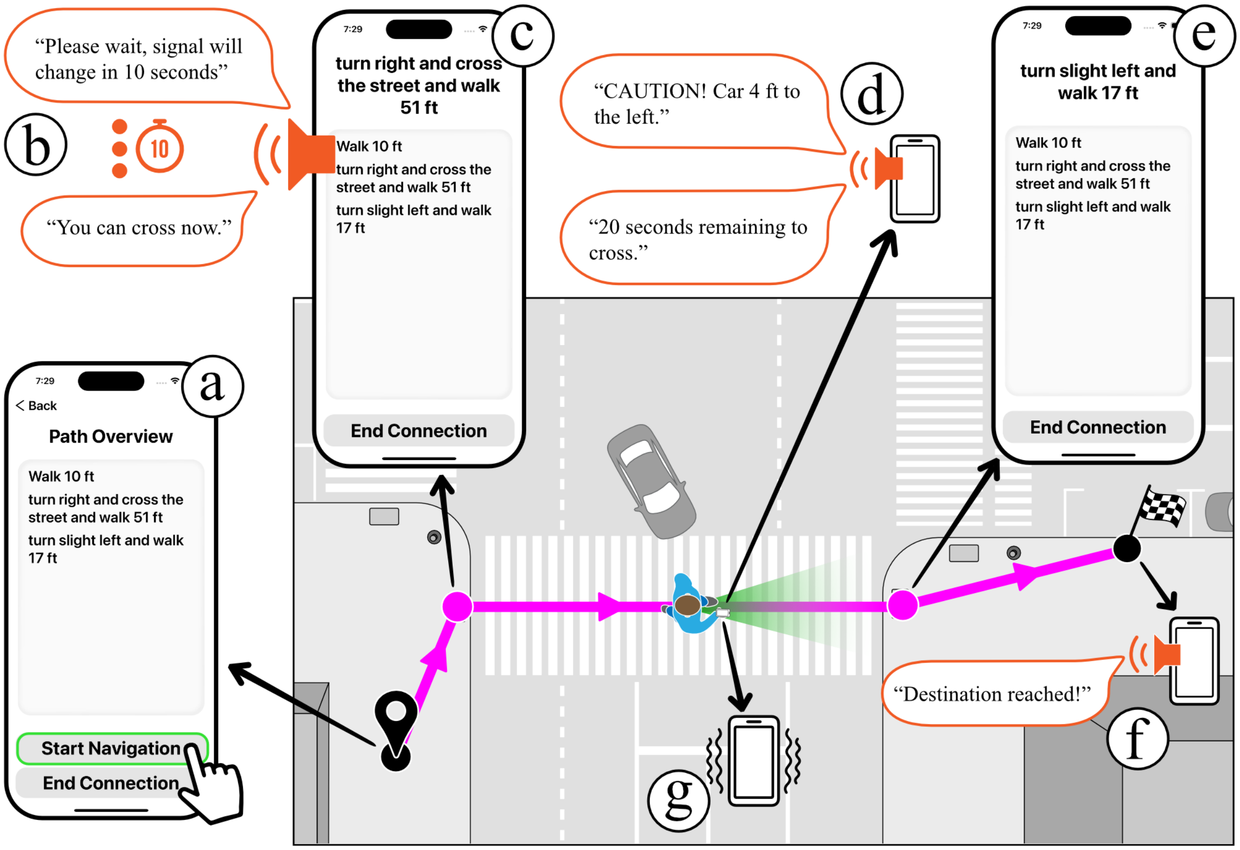
Figure 3.6: Recognizing pedestrian signal states from the camera's video feed. StreetNav compares the number of white and red pixels in the signal crops to determine its state: (a) *walk* vs. (b) *wait*.

first initiate a connection request through the app, which activates the gesture-based localization (Section 3.3.2) in the computer vision pipeline. The app prompts the user to wave one hand over their head (Figure 3.3b), enabling the system to begin tracking their precise location on the map (Figure 3.3d). BLV users can then select a destination from nearby POIs and begin receiving navigation feedback through the app.

Figure 3.7 shows the StreetNav app's user interface, which uses audiohaptic cues for (i) providing routing instructions, (ii) preventing veering off track, (iii) notifying about nearby obstacles, and (iv) assisting with crossing streets. Upon reaching the destination, the app confirms their arrival. The following sections describe the app's interface in detail.

Providing routing instructions

The app conveys routing instructions to the users by first giving an overview of the route and then announcing each instruction, in situ, based on their current location in the environment. Figure 3.7a shows the app screen with the path overview. Prior work on understanding BLV people's navigation behaviors [151, 152, 153] reveals that BLV people often prepare for their routes before actually walking through them. StreetNav assists them in this preparation by giving an overview of the path before beginning navigation. The path overview consists of several instructions, with each helping them get from one waypoint to the next. BLV users read through the path overview using VoiceOver [141]. Users then tap the 'Start Navigation' button, which announces each instruction when they reach a waypoint. Figure 3.7b–f shows how the app dynamically updates the


Path Overview

Walk 10 ft

turn right and cross the street and walk 51 ft

turn slight left and walk 17 ft

turn right and cross the street and walk 51 ft

Walk 10 ft

turn right and cross the street and walk 51 ft

turn slight left and walk 17 ft

turn slight left and walk 17 ft

Walk 10 ft

turn right and cross the street and walk 51 ft

turn slight left and walk 17 ft

a
b
c
d
e
f
g
7:29
< Back
Start Navigation
End Connection
“Please wait, signal will change in 10 seconds”
10
“You can cross now.”
“CAUTION! Car 4 ft to the left.”
“20 seconds remaining to cross.”
“Destination reached!”

Figure 3.7: The StreetNav App’s user interface. It provides routing instructions to their destination via (a) a path overview and (c, e) real-time feedback that updates their current instruction based on their location. Upon reaching a sidewalk, (b) the app informs the user about when it is safe to cross and (d) how much remains for them to cross over. It also (d) notifies the user of a nearby obstacle’s category and relative location to help them avoid it. The app (f) confirms the user’s arrival at the destination. Throughout the journey, the app provides (g) continuous audiohaptic feedback to prevent users from veering off track.

next instruction based on the user’s location in the environment. Throughout the journey, users can access the path overview and the current navigation instructions on demand via VoiceOver.

Preventing veering off track

Figure 3.8 illustrates the app’s feedback for preventing users from veering off track. Given the user’s current position, heading, and destination route, StreetNav computes the *direction* and *extent* of veering. To convey *direction* of veering, we used 3D spatialized sound, which plays continuous beeping sounds from the right speaker when users veer to the left (Figure 3.8a) and from the left

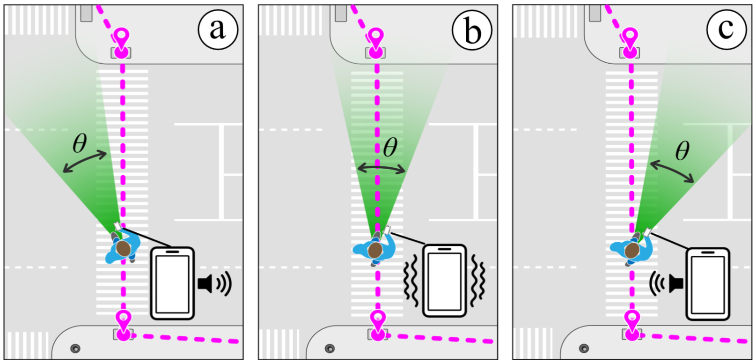


Figure 3.8: Audiohaptic cues for preventing users from veering off track. Sample user trajectories showing feedback when users veer (a) left, (b) do not veer, and (c) veer to the right. When the user's heading coincides with the route to the destination, within a tolerance angle θ (highlighted in green), users receive (b) subtle haptic vibrations to reinforce them. When they veer off the route, outside the tolerance angle θ, they hear spatialized beeping sounds that are rendered from the (a) right speaker when veering left, and from the (c) left speaker when veering right.

speaker when users veer to the right (Figure 3.8c). Users can follow the direction of the beeping sound to correct for veering. To convey the *extent* of veering. i.e., how severely the user is veering, we render the frequency of beeps to be proportional to the angle between the user's current heading and the route. As users veer away from the correct direction, the frequency of beeps increases; and when they begin to turn towards the correct direction, the frequency of beeps decreases. Users can also leverage the frequency of beeps to determine how to correct for veering, by always moving in the direction where the beeps' frequency reduces. This enables users to correct for veering even without the spatialized sound feedback we used for direction. This eliminates the need to wear headphones to understand spatialized sound.

We ran pilot experiments to test this feedback mechanism with our BLV co-author. We found that the continuous audio feedback was helpful but also became overwhelming as it forced them to strictly follow StreetNav's route. To address this, we relaxed the veering requirements by introducing a tolerance angle (θ). Figure 3.8 shows the tolerance angle in green color, which is depicted as a cone centered at the user's current heading. We updated the veering feedback to only play beeping sounds when users veer off in either direction by at least $\theta/2$ degrees. To maintain the continuity of feedback, we chose to render subtle haptic vibrations when users move in the correct direction within the tolerance angle. Within this tolerance angle, the intensity of the haptic vibration increases when users approach the exact correct heading and decreases when they start to veer off. This is similar to how the frequency of beeps increases when users veer away. In this way, the audio feedback acts as negative reinforcement, and the haptic feedback acts as positive reinforcement. Figure 3.8b illustrates the haptic feedback. We experimentally tuned the tolerance angle, θ, and set its value for our system to 50°.

To generate the audiohaptic cues, the app receives the user's current position and destination route from the computer vision pipeline. For the user's current heading, we experimented using the user's trajectory to predict their heading using the Kalman filter. This approach, however, yields inaccurate headings due to the noisy tracking data. Thus, we leveraged the smartphone's compass to determine the user's current heading. We offset the compass readings by a fixed value to ensure that its zero coincides with the map's horizontal direction. This enabled us to perform all heading-related computations with respect to the map's frame of reference.

Notifying about nearby obstacles

Figure 3.7d shows how StreetNav alerts the user of obstacles nearby. The app announces the obstacle's category, distance, and relative location. For example, when a car approaches the user, the app announces: "*Caution! Car, 4 ft. to the left.*" Similar to veering feedback, the relative location is computed using both the computer vision pipeline's outputs and the smartphone's compass reading.

We tried feedback formats with varying granularity to convey the obstacle's relative location. First, we experimented with *clock-faced directions*: "*Car, 4 ft. at 1 o'clock.*" Clock-faced directions are commonly used in many GPS-based systems such as BlindSquare to convey directions. We learned from pilot evaluations with our BLV co-author that this feedback format was too fine-grained, as it took them a few seconds to decode the obstacle's location. This does not fare well with moving obstacles, such as pedestrians, that may have already passed the user before they are able to decode the location. Moreover, StreetNav's goal with obstacle awareness is to give users a quick idea that something is nearby them, which they can then use to circumnavigate via their mobility skills. To address this, we tried the more coarse format with just four directions: left, right, front, and back. This was found to give users a quick intimation, compared to the clock-faced directions.

Assisting with crossing streets

The StreetNav app helps users cross streets by informing them *when* to cross and how much time remains before the signal changes.

Figure 3.7b and Figure 3.7d illustrate the feedback. Upon reaching a sidewalk corner, the app checks for the signal state recognized by the computer vision pipeline. If the signal is '*wait*' when the user arrives, the app informs the user to wait along with the time remaining before the signal changes. If the signal is '*walk*' when the user arrives, the app informs the user to begin crossing only if the time remaining is sufficient for crossing. For the intersection used in our user studies, this was experimentally found to be 15 seconds. Otherwise, the user is advised to wait for the next cycle. Once the user begins crossing on the '*walk*' signal, the app announces the time remaining for them to cross over. This feedback is repeated at fixed intervals until the user reaches the other sidewalk corner. We experimentally fine-tuned this interval with feedback from our BLV co-author. We tried several intervals, such as 5, 10, and 15 seconds, and found that shorter intervals overwhelmed the users, whereas longer intervals practically would not be repeated enough times to give them meaningful information. We settled on repeating the feedback every 10 seconds for our implementation.

3.4 User Study

Our user study had three goals, related to RQ2 and RQ3. First, we wanted to evaluate the extent to which StreetNav addressed BLV pedestrians' challenges in navigating outdoor environments when using existing GPS-based systems. Through our formative interviews (Section 3.2), we discovered three main challenges: routing through complex environment layouts (**C1**), avoiding unexpected obstacles (**C2**), and crossing street intersections (**C3**). Second, we wanted to an-

Table 3.2: Self-reported demographics of our study participants. Gender information was collected as a free response. Participants rated their familiarity with assistive technology (AT) on a scale of 1–5.

PID	Age	Gender	Occupation	Race	Vision ability	Onset	Mobility aid	AT familiarity (1–5)
P1	24	Male	App developer	Asian	Low vision	Age 19	White cane	2: Slightly familiar
P2	28	Male	Data manager	White	Low vision	At birth	None	3: Moderately familiar
P3	48	Male	Not employed	Black	Totally blind	Age 32	White cane	3: Moderately familiar
P4	46	Female	Social worker	Latino	Totally blind	Age 40	White cane	4: Very familiar
P5	43	Female	Not employed	Asian	Totally blind	At birth	White cane	4: Very familiar
P6	52	Male	Mgmt. analyst	Mixed	Light perception only	Age 9	White cane	5: Extremely familiar
P7	26	Female	Writer	Mixed	Low vision	At birth	White cane	2: Slightly familiar
P8	51	Male	Not employed	Black	Light perception only	Age 26	Guide dog	3: Moderately familiar

alyze BLV pedestrians' experience of navigating outdoors using StreetNav compared to existing GPS-based systems. Third, we wanted to see how participants rank the two navigation systems—StreetNav vs. GPS-based system—in order of their preference for outdoor navigation assistance.

3.4.1 Study Description

Participants

We recruited eight BLV participants (five males, three females; aged 24–52) by posting to social media platforms and by snowball sampling [94]. Participants identified themselves with a range of racial identities (Asian, Black, White, Latino, and Mixed) and all of them lived in a major city in the US. Participants also had diverse visual abilities, onset of vision impairment, and familiarity with assistive technology (AT) for navigation.

Table 3.2 summarizes participants' information. All but three participants (P1, P7, and P8) reported themselves as being moderately–extremely experienced with AT for navigation (3+ scores on a 5-point rating scale). Only P3 reported minor hearing loss in both ears and wore hearing aids. All participants except two (P2, P9) used white cane as their primary mobility aid. P2 did not use any mobility aid, while P9 primarily used a guide dog for navigation. The IRB-approved study lasted for about 120 minutes, and participants were compensated $75 for their time.

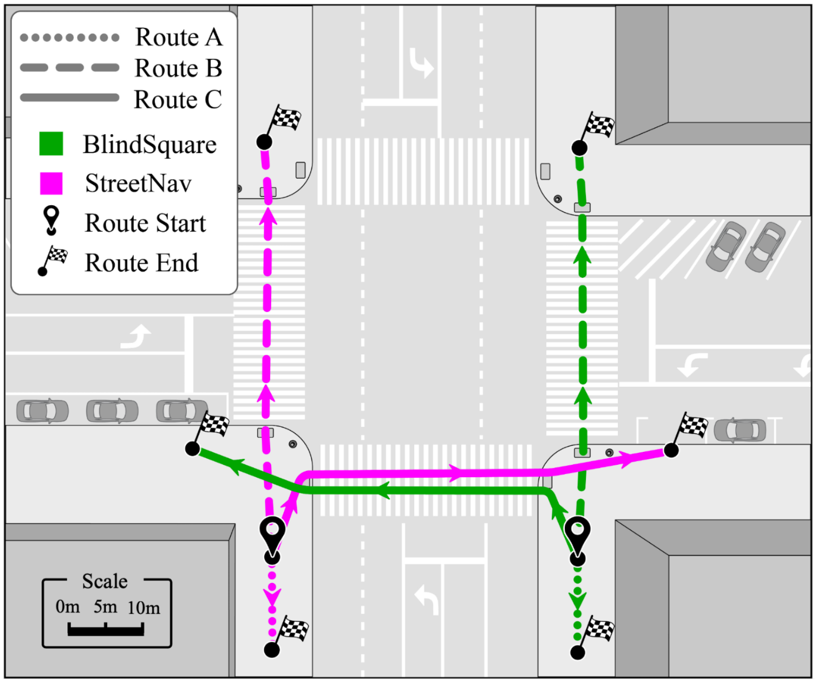


Figure 3.9: The routes used in the navigation tasks. (A) 12 meters, stationary person to avoid on the sidewalk. (B) 30 meters, cross street, and moving person to avoid on the sidewalk. (C) 38 meters, a 90° turn, cross street, and moving person to avoid on the crosswalk. To mitigate learning effects, routes for the two conditions are symmetrically designed, situated on opposite sides of the street.

Experimental Design

In the study, participants completed three navigation tasks at a street intersection in two conditions: (i) StreetNav and (ii) BlindSquare [97], a popular GPS-based navigation app especially designed for BLV people. We evaluated the two systems via their respective iOS apps on an iPhone 14 Pro. Both systems' apps seamlessly integrated with VoiceOver, and all eight participants had a high level of familiarity with using iPhones and VoiceOver, with ratings of 3 or higher on a 5-point scale. During the study, participants continued to use their primary mobility aids, such as white canes and guide dogs, in both conditions. This approach allowed us to make a meaningful comparison between StreetNav and the BLV pedestrians' current methods of outdoor navigation, simulating their usual practice of incorporating GPS-based navigation systems alongside their mobility aids.

Our study followed a within-subjects design, in which participants tested the two navigation systems in a counter-balanced order to minimize potential order-bias and learning effects. In each condition, participants were tasked with completing three distinct navigation challenges, corresponding to three specific routes. Figure 3.9 illustrates these three navigation routes. We deliberately chose the routes to lie within the street camera's field of view and include a range of difficuly levels for each task: (A) a short route, 12 meters, that involved avoiding a stationary person on the sidewalk, (B) a long route, 30 meters, that involved crossing a street and avoiding a moving person on the sidewalk, and (C) a complex route, 38 meters, that involved making a 90 degree turn, crossing a street, and avoiding a moving person on the crosswalk. For each of these tasks, one of our researchers assumed the role of the obstacle. Notably, none of the participants were familiar with the specific street intersection selected as the study's location.

Given that participants navigated the same intersection in both conditions, the potential for learning effects as a confounding factor was carefully considered. To address this concern, we took deliberate measures by creating distinct routes for each condition. Specifically, we designed the routes in both conditions to be symmetric—rather than being identical—with the starting and ending points of each route strategically positioned on opposite sides of the street intersection, as illustrated in Figure 3.9. The symmetry of routes ensured that participants encountered the same challenges in both conditions. To ensure participants' safety, the researchers accompanied them at all times during the study, prepared to intervene whenever necessary.

Procedure

We began each study condition by giving a short tutorial of the respective smartphone app for the system. During these tutorials, participants were taught how to use the app and how to interpret the various audiohaptic cues it offered. To accommodate potential challenges arising from ambient noise at the street intersection, participants were given the option to wear headphones during the study. Only two participants, namely P3 and P5, exercised that option; rest of the participants relied on the smartphone's built-in speaker to hear the audiohaptic cues.

After completing the three navigation tasks for each condition, we administered a questionnaire comprising four distinct parts. These parts were designed to assess participants' experiences around challenges faced by BLV pedestrians in outdoor navigation, specifically addressing the following aspects: routing to destination (**C1**), veering off course (**C1**), avoiding obstacles (**C2**), and crossing streets (**C3**). It included questions about how well each system assisted with the challenges, if at all. Participants rated their experience on a 5-point rating scale, where a rating of "1" indicated "*not at all well,*" and a rating of "5" indicated "*extremely well.*" After each part of the questionnaire, we asked follow-up questions to gain deeper insights into the reasons behind their ratings and their overall experiences.

Following their experience with both navigation systems, participants were asked to complete a post-study questionnaire. This questionnaire required them to rank the two navigation systems in terms of their preference for outdoor navigation. Subsequently, we directed our discussion toward StreetNav, engaging participants in a conversation about potential avenues for improvement. We also inquired about the specific scenarios in which they envision using this system in the future.

In addition to the questionnaires that aimed at capturing participants' subjective experiences, we also gathered system usage logs and video recordings of participants throughout the study. These objective data sources, including usage logs and video recordings, allowed us to perform a comprehensive analysis of participants' actual performance in the navigation tasks. It is worth noting that willingness to be video-recorded was completely voluntary, i.e., did not affect participants' eligibility or compensation. All eight participants still agreed to be video-recorded, providing us with written consent to do so.

Analysis

We report participants' spontaneous comments that best represent their overall opinions, providing further context on the quantitative data we collected during the study. We analyzed the transcripts for participants' quotes and grouped them according to the (i) questionnaire's four parts: routing to destination, veering off course, avoiding obstacles, and crossing streets; (ii) overall satisfaction and ranking preferences, and (iii) how users' individual experiences influenced their preferences.

3.4.2 Results

Our results reveal that StreetNav helped participants reach their destinations with more precision, gain awareness of obstacles, reduce veering off course, and confidently cross streets. For the statistic analysis of each measure, we first conducted a Kolmogorov-Smirnov test to determine if the data was parametric or non-parametric. Then, when comparing between the two conditions, we used a paired t-test when the data was parametric. In addition to quantitative measures, we conducted a detailed analysis of video recordings, manually annotating the routes participants took during the study. We provide these metrics to offer additional insights into participants' performance across both experimental conditions.

Routing to Destination

Figure 3.10 shows participants' average rating for their experience following routes to the destination in each condition. The mean (± std. dev.) rating for participants' perceived usefulness of the routing instructions in guiding them to the destination was 4.13 (±0.64) for StreetNav and 2.38 (±0.91) for BlindSquare. The condition had a significant main effect ($p = 0.014$) on participants' experience reaching destinations with the routing instructions. The mean (± std. dev.) rating for participants' experience with the system's ability to track them was 4.50 (±0.76) for StreetNav and 2.88 (±1.13) for BlindSquare. The condition had a significant main effect ($p = 0.001$) on participants' perception of how well the system tracked them en route to the destination. This indicates that participants found StreetNav more useful than BlindSquare for guiding them to the destination.

Figure 3.11 illustrates our analysis of the video recordings, plotting the typical paths taken by participants in the third route across both conditions. We computed various metrics from their paths, that provide insights into participants' self-reported ratings.

We found that when using BlindSquare, participants covered greater distances to reach the same destinations compared to when using StreetNav. On average, participants traveled a distance approximately 2.1 times longer than the shortest route when relying on BlindSquare. In contrast, when using StreetNav, they covered a distance of only about 1.1 times the shortest route to their destination. This represents a 51% reduction in the unnecessary distance traveled with StreetNav in comparison to BlindSquare. Figure 3.11b shows how participants using BlindSquare often exhibited an oscillatory pattern near their destinations (P1, P8) before eventually reaching close to them.

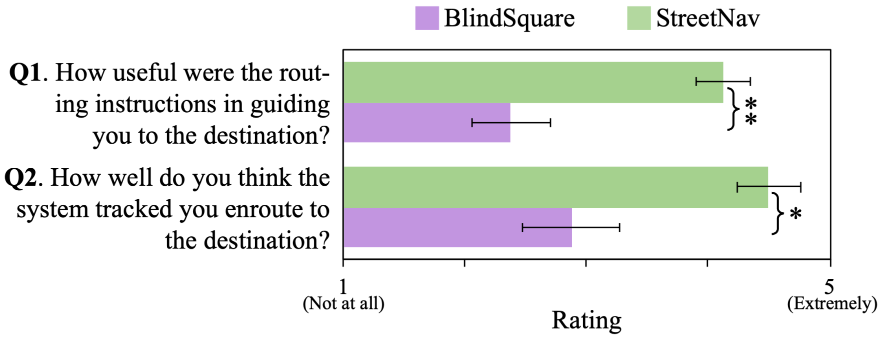
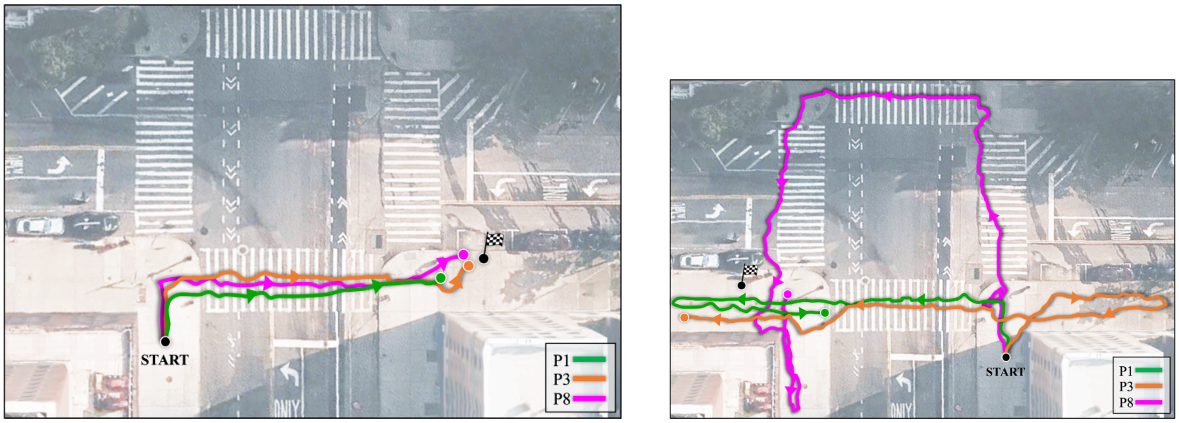


Figure 3.10: Results for participants' experience with routing to the destination. Participants rated the (1) usefulness of routing instructions, and (2) the system's ability to track them en route to the destination. Participants found StreetNav'sturn-by-turn instructions significantly more useful and precise than BlindSquare's "as the crow flies"-style routing instructions. Pairwise significance is depicted for $p < 0.01$ ($*$) and $p < 0.05$ ($**$). The error bars indicate standard error.

Additionally, StreetNav's routing instructions displayed a notably higher level of precision, guiding participants to their destinations with 2.9 times greater accuracy than BlindSquare. Figure 3.11 clearly shows this trend for the third route. On average, across the three study routes, participants using StreetNav concluded their journeys within a tighter radius of 12.53 feet from their intended destination. In contrast, participants relying on BlindSquare concluded their journeys within a radius of 35.94 feet from their intended destination. Two study participants, P4 and P5, even refused to navigate to the destination in two of the three tasks with BlindSquare. This was primarily attributed to BlindSquare's low precision in tracking the participants and often guiding them to take incorrect turns. Figure 3.11b highlights how BlindSquare caused P8 to go around the intersection before finally getting close the destination.

Participants preferred StreetNav over BlindSquare for its audiohaptic cues for turn-by-turn



(a) StreetNav

(b) BlindSquare

Figure 3.11: Comparison of paths traveled by three participants (P1, P3, P8) for route 'C' using (a) StreetNav, and (b) BlindSquare. StreetNav's routing instructions consistently guided participants to the destination via the shortest path. BlindSquare, however, caused participants to take incorrect turns (P1, P3, P8), oscillate back and forth near destinations (P1, P8), and even go around the whole intersection before getting close to the destination (P8).

navigation instructions, which they found to be more useful and precise than BlindSquare's "as the crow flies"-style clock face and distance-based instructions. P3's comment encapsulates this sentiment:

> "*When it's time for me to turn right and walk a certain distance, [StreetNav] is very, very, very precise.*" **–P3**

Although all participants preferred StreetNav's routing feedback over BlindSquare's, distinct patterns emerged in their preference and utilization of these cues. StreetNav delivers a combination of audiohaptic and speech feedback for routing, and participants adopted varying strategies for utilizing this feedback. Some individuals placed greater reliance on the veering haptic feedback as their primary directional guide, while reserving speech feedback as a fallback option. Conversely, some participants prioritized the speech feedback, assigning it a higher level of importance in their navigation process compared to audio-haptic cues.

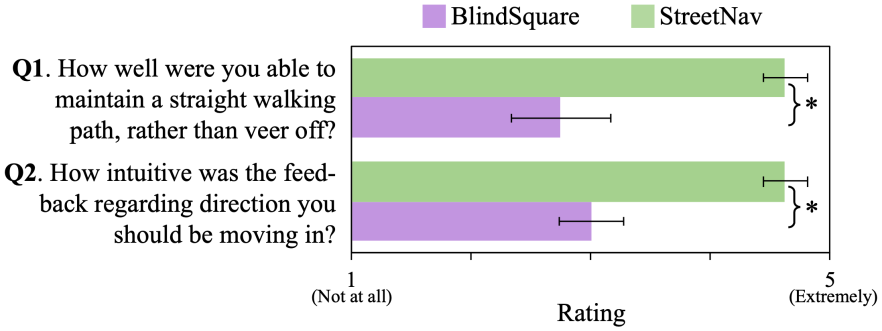


Figure 3.12: Results for participants' perceived ability to prevent veering off path. Participants rated their ability to (1) maintain a straight walking path, and (2) intuitiveness of the feedback regarding direction they should be moving in; on a scale of 1–5. StreetNav'saudiohaptic feedback was significantly more intuitive than BlindSquare's in preventing participants from veer off path. Pairwise significance is depicted for $p < 0.01$ ($*$). The error bars indicate standard error.

Veering Prevention

Figure 3.12 shows participants' average rating for their perceived ability to (1) maintain a straight walking path, i.e., prevent veering off course, and (2) intuitiveness of the feedback they received regarding direction to move in. The mean (± std. dev.) rating of participants' perceived ability to maintain a straight walking path with StreetNav was 4.63 (±0.52) and with BlindSquare was 2.75 (±1.17). The condition had a significant main effect ($p = 0.001$) on participants' perceived ability to prevent veering off course. The mean (± std. dev.) rating for intuitiveness of the feedback that helped them know which direction to move in was 4.63 (±0.52) for StreetNav and 3.00 (±0.76) for BlindSquare. The condition had a significant main effect ($p = 0.006$) on intuitiveness of feedback that helped participants prevent veering off path.

Our examination of the video recordings aligns closely with participants' ratings. It reveals that StreetNav minimized participants' deviations from the shortest path to the destinations in comparison to BlindSquare. Over the course of the three routes, participants displayed an average deviation from shortest path, that was reduced by 53% when using StreetNav as opposed to BlindSquare.

With BlindSquare, many participants reported difficulty maintaining awareness of their surroundings, including both obstacles and navigation direction, which frequently led to deviations from their intended paths. For instance, P2 reported challenges in maintaining their orientation with the need to avoid obstacles:

> *"[BlindSquare] basically demanded me to keep track of my orientation as I was moving, which is pretty difficult to do when you're also trying to keep other things in mind, like not bumping into things."* **–P6**

In contrast, StreetNav effectively addressed this challenge by providing continuous audiohaptic feedback for maintaining a straight walking path, instilling a sense of confidence in participants. P3, who tested StreetNav before BlindSquare, reflected on their desire for a similar continuous feedback mechanism within BlindSquare, akin to the experience they had with StreetNav:

> *"[with BlindSquare] even though I couldn't see the phone screen, my eyes actually went towards where I'm holding the screen. It is almost as if on a subconscious level, I was trying to get more feedback. With [StreetNav] I had enough feedback."* **–P3**

Many participants appreciated StreetNav's choice of haptic feedback for veering. Some participants envisioned the haptic feedback to be especially useful in environments with complex layouts:

> "*In the [areas] where the streets are very slanted and confusing. I think haptic feedback will be especially helpful."* **–P5**

Other participants highlighted the advantage of haptic feedback in noisy environments where audio and speech feedback might be less effective.

However, both P4 and P6 exclaimed that StreetNav's haptic feedback would only work well when holding the phone in their hands. This meant that hands-free operation of the app may not be possible, which is important for BLV people since one of their hands is always occupied by the white cane. P4 proposed integrating the app with their smartwatch for rendering the haptic feedback to enable hands-free operation.

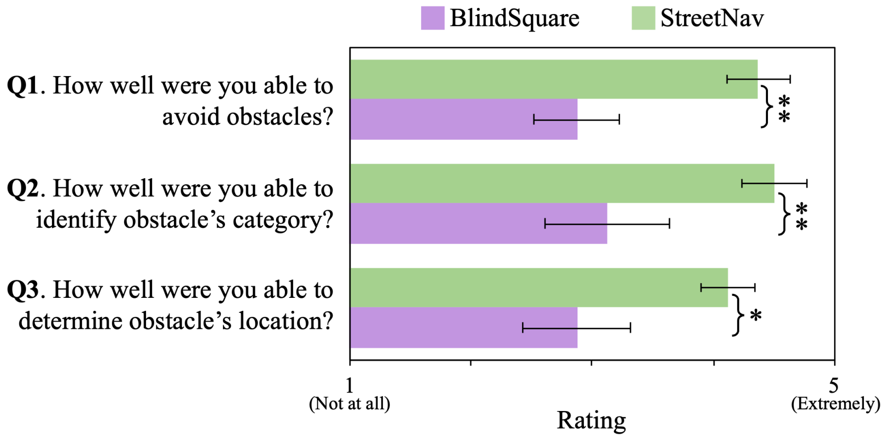


Figure 3.13: Results for participants' perceived obstacle awareness. Participants rated their ability to (1) avoid obstacles, (2) identify its category (e.g., person, bicycle), and (3) determine its relative location; on a scale of 1–5. StreetNav significantly improved participants' awareness of nearby obstacles during navigation. Pairwise significance is depicted for $p < 0.01$ ($*$) and $p < 0.05$ ($**$). The error bars indicate standard error.

Obstacle Awareness

Figure 3.13 shows participants' average rating for their perceived awareness of obstacles across the two conditions. Specifically, participants rated their ability to (1) avoid obstacles, (2) identify its category (e.g., person, bicycle, trash can), and (3) determine its relative location. The mean (± std. dev.) rating for participants' perceived ability to avoid obstacles was 4.38 (±0.74) for StreetNav and 2.88 (±0.99) for BlindSquare, to identify its category was 4.50 (±0.76) for StreetNav and 3.13 (±1.46) for BlindSquare, and to determine obstacle's relative location was 4.13 (±0.64) for StreetNav and 2.88 (±1.25) for BlindSquare. A paired t-test revealed that the condition had a significant main effect on participants' perceived ability to avoid obstacles ($p = 0.030$), identify its category ($p = 0.037$), and relative location ($p = 0.004$). This suggests that StreetNav offered users a heightened awareness of nearby obstacles compared to the baseline condition of BlindSquare.

With StreetNav, participants had the option to use obstacle avoidance audio feedback in conjunction with their conventional mobility aids. However, in the case of BlindSquare, the system itself did not offer any obstacle-related information. Consequently, participants primarily relied on their traditional mobility aids in this condition, as is typical when using GPS-based systems. Our analysis of the video recordings found that in both experimental conditions, participants encountered no instances of being severely hindered by obstacles. Instead, they adeptly navigated around obstacles with the assistance of their white canes or guide dogs.

Although participants generally had a positive perception of obstacle avoidance when using StreetNav, their opinions on the utility of obstacle awareness information varied. Some participants found this information beneficial, emphasizing its role in preventing "*awkward bumping into people*" (**P2**) and boosting their confidence, resulting in greater *"speed in terms of walking"* (**P3**). Conversely, participants who felt confident avoiding obstacles with their mobility aids regarded StreetNav's obstacle information to be extraneous. P8 also expressed concerns about the potential information overload it could cause in dense urban areas:

> *"To know where people are, is a bit of overkill. If you turn this thing on in New York City, it would have your head go upside down."* –**P8**

Many participants proposed an alternative use case for StreetNav's obstacle awareness information, highlighting its potential for providing insights into their surroundings. They suggested that this information could unlock environmental affordances, including the identification of accessible light signals and available benches for resting: "*knowing there was a bench was top-notch for me*" (**P8**). Therefore, StreetNav's obstacle awareness information served a dual purpose, aiding in both obstacle avoidance and environmental awareness, allowing users to "*know what's around*"(**P8**) them.

Crossing Streets

Figure 3.14 shows participants' average rating for their perceived comfort in crossing streets. The mean (± std. dev.) rating of participants' perceived comfort in making the decision on when to begin crossing the street was 4.50 (±0.76) for StreetNav and 2.88 (±1.64) for BlindSquare. The mean (± std. dev.) rating of participants' perceived comfort in safely making it through the crosswalk and reach the other end was 4.63 (±0.52) for StreetNav and 2.00 (±1.41) for BlindSquare. A paired t-test showed that the condition had a significant main effect on participants' comfort in beginning to cross streets (p = 0.029) and in safely making it to the other side (p = 0.001).

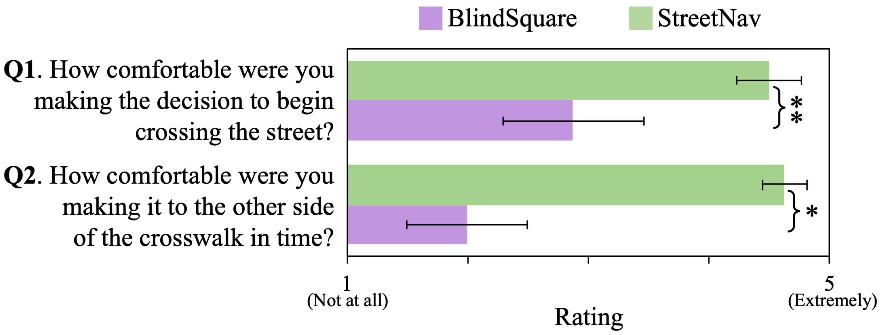


Figure 3.14: Results for participants' perceived comfort in crossing streets. Participants rated their perceived comfort in (1) making the decision on when to begin crossing the street, and in (2) pacing themselves when crossing. Participants were significantly more comfortable crossing streets with StreetNav in comparison to BlindSquare. Pairwise significance is depicted for $p < 0.01$ ($*$) and $p < 0.05$ ($**$). The error bars indicate standard error.

As BlindSquare does not provide feedback specifically for crossing streets, participants reported relying on their auditory senses, listening for the surge of parallel traffic. However, during the semi-structured interviews, some participants highlighted challenging scenarios that can make this strategy less reliable. P4, for instance, pointed out that ironically, less traffic can complicate street crossings:

> *"I don't always know when to cross because it's so quiet. And sometimes two, three light cycles go by, and I'm just standing there."* **–P4**

This issue has been exacerbated by the presence of electric cars, which are difficult to hear due to their quiet motors. For P3, their hearing impairments made it challenging to listen for traffic. Thus, most participants appreciated StreetNav's ability to assist with crossing streets:

> "*When it's quiet, I would cross. But now with hybrid cars, it's not safe to do that either. The [StreetNav] app telling you which street light is coming on is really helpful.*" **–P7**

Participants made decisions to cross the streets by combining StreetNav's feedback with their auditory senses. Many participants emphasized that having information about the time remaining to cross significantly boosted their confidence, especially when this information aligned with the sounds of traffic: "*I thought it was great because I could tell that it matched up*" (**P8**). This alignment between the provided information and their sensory perception inspired confidence in participants:

> "*Relying on my senses alone feels like a gamble about 90 percent of the time, so a system like [StreetNav] that accurately displays the amount of time I have to cross the street is great.*" **–P2**

3.4.3 Forced Ranking Results

All eight participants unanimously chose StreetNav over BlindSquare as their preferred navigation assistance system. We asked participants to also rank their preferred type of routing instructions. All eight participants strongly preferred StreetNav's turn-by-turn routing instructions compared to BlindSquare's "as the crow flies," direction and distance-style routing instructions.

In the semi-structured interview, participants were asked to elaborate on their rankings. Participants pointed out multiple navigation gaps in BlindSquare, with P2 summarizing participants' sentiment:

> "*If you're only getting somebody 90 percent of the way there, you're not really achieving what I would consider to be the prime functionality of the system.*" **–P2**

In contrast, participants praised StreetNav for its precision and real-time feedback, emphasizing the importance of granular and holistic information to support all facets of navigation. However, participants did acknowledge occasional "glitchiness" (**P7**) with StreetNav, which occurred when they moved out of the camera's field of view or were occluded by other pedestrians or vehicles, resulting in lost tracking. Nevertheless, participants still regarded StreetNav as a significant enhancement to their typical navigation experiences, expressing increased confidence in exploring unfamiliar outdoor environments in the future.

> "*It would encourage me to do things that I would not usually... It would make me more confident about going out by myself.*" **–P4**

Participants also appreciated StreetNav's ability to identify them in near real-time:

> "*What I found very interesting about the connection part is how quickly it identifies where I am, as soon as I waved my hand, it senses me.*" **–P3**

Participants also provided suggestions for improving StreetNav. Some participants wanted a hands-free version that would allow them to hold a white cane in one hand while keeping the other free. Additionally, while they found the gesture of waving hands for connecting with the system socially acceptable, they acknowledged that it might be perceived as somewhat awkward by others in the street.

> "*[Waving a hand] may seem kind of weird to people who don't understand what is going on. But for me personally, I have no issue.*" **–P3**

While the gesture-based localization was generally accurate, there were instances where other pedestrians were incorrectly detected as the study participant. On average, the gesture-based localization worked accurately over 90% of the time.

3.4.4 How Individual Experiences Influenced Participants' Preferences

Throughout the study, participants offered feedback based on their unique backgrounds. We observed distinct patterns in their preferences, affected by their (i) onset of vision impairment, (ii) level of vision impairment, and (iii) familiarity with assistive technology.

Onset of vision impairment

Participants with early onset blindness preferred nuanced, concise feedback with an emphasis on environmental awareness. They used the system as an additional data point without complete reliance. In contrast, participants with late onset blindness trusted the system more and relied heavily on its feedback.

Level of vision impairment

Totally blind participants appreciated the veering feedback, while low-vision users, who had more visual information, relied on their senses and did not need as much assistance with veering.

Low-vision participants appreciated the street crossing feedback rather than trying to glean information from pedestrian signals across the street. Totally blind participants relied more on listening for parallel traffic—their usual mode of operation—and used StreetNav's street-crossing feedback as a confirmation.

Familiarity with assistive technology (AT)

We noticed that participants who commonly use AT for navigation quickly adapted to StreetNav, while those with less experience hesitated in trusting StreetNav's feedback and had a slightly steeper learning curve. Still, all participants mentioned feeling more comfortable with StreetNav as the study progressed. Both groups also expressed increased confidence in exploring new areas with StreetNav.

3.5 Discussion

Our goal with StreetNav was to explore the idea of repurposing street cameras to support precise outdoor navigation for BLV pedestrians. We reflect upon our findings to discuss how street camera-based systems might be deployed at scale, implications of a street camera-based navigation approach for existing GPS-based navigation systems, and the affordances enabled by precise, real-time outdoor navigation assistance.

Deploying street camera-based navigation systems at scale

StreetNav demonstrates that street cameras have the potential to be repurposed for supporting precise outdoor navigation for BLV pedestrians. Our study results show that street camera-based navigation systems can guide users to their destination more precisely and prevent them from veering off course (Figure 3.11). Our results also show that street camera-based systems can support real-time, scene-aware assistance by notifying users of nearby obstacles (Figure 3.13) and giving information about when to cross streets ((Figure 3.14)). These benefits of a street camera-based approach, over existing GPS-based systems, underscores the need for deploying such systems at scale. Although our system, StreetNav, was deployed at a single intersection, we learned insights on potential challenges and considerations that must be addressed to deploy street camera-based systems at scale.

Several internal and external factors need to be considered before street cameras can be effectively leveraged to support blind navigation at scale. External factors, including lighting conditions and occlusions on the street, may affect system performance. For instance, we noticed that StreetNav's ability to track pedestrians was affected severely in low-light conditions (e.g., at night) and by occlusions due to the presence of large vehicles (e.g., trucks, buses) and the installation of scaffoldings for construction. Such challenges affect the reliability of street camera-based systems and may limit its operational hours. Internal factors, including the positioning of cameras, their field of view, and variability in resolution, may affect the extent to which such systems can promise precise navigation assistance. For instance, the visibility of the pedestrian signals from the camera feed could affect how much such systems can assist users with crossing streets. With StreetNav, we observed a drop in tracking accuracy as individuals and objects moved further away from the camera.

Therefore, deploying street camera-based systems at scale would require future work to investigate the extent to which both external factors (e.g., lighting, occlusions) and internal factors (e.g., camera resolution) affect system performance and reliability.

To address some of the technical limitations around tracking performance and field of view limitations, future research could explore integrating multiple cameras at various elevations and viewing angles. Prior work on robot navigation has explored the fusion of multiple cameras to improve tracking performance [132, 133, 134]. Future work could also explore an ecosystem of accessible street cameras that can share information to automatically manage hand-offs across street intersections, providing users with a seamless experience beyond a single street intersection. Such ecosystems, which span beyond one intersection to a whole district or city, could enable new affordances, such as automatically sensing pedestrian traffic to inform traffic signals and vice versa.

Implications for GPS-based navigation systems

When cameras are available, and conditions align favorably, street camera-based systems offer BLV individuals a valuable source of fine-grained, high-precision information, significantly enhancing their navigational experience and environmental awareness. These capabilities are currently beyond the reach of conventional GPS-based systems. All eight study participants unanimously chose StreetNav over BlindSquare as their preferred navigation system due to its precise, scene-aware navigation assistance (Section 3.4.3). However, it's important to acknowledge that street camera-based systems have their own set of limitations. The widespread availability of street cameras is not yet a reality, and ideal conditions may not always be met for their effective use. In contrast, GPS-based systems, while lacking in precision and environmental awareness, are universally accessible and resilient in varying conditions, including low light. A harmonious integration of these two approaches is a promising solution. Users can tap into street-camera information when conditions permit, seamlessly transitioning to GPS data when necessary. This can be facilitated through sensor fusion or information hand-offs, creating a synergy that ensures a smooth and reliable navigational experience. Future approaches could explore how these two systems can effectively complement each other, addressing their respective limitations and enhancing overall performance.

Affordances of precise outdoor navigation assistance for BLV people

Previous research in indoor navigation has demonstrated the advantages of accurately pinpointing users' locations [114, 110, 113] and providing scene-aware navigational information [148, 150]. However, achieving such precision has remained a challenge in outdoor environments, primarily due to the limited accuracy of GPS technology [103]. StreetNav's approach of leveraging existing street cameras demonstrates that precise outdoor navigation support for BLV pedestrians is possible. Our study reveals the advantages of precise, fine-grained navigation for BLV individuals. These benefits include a substantial reduction in instances of veering and routing errors, such as deviation from the shortest path or missing intended destinations, as well as augmented environmental awareness.

StreetNav offered our participants a glimpse into the potential of precise outdoor navigation. Several participants desired even greater precision, including the ability to discern the exact number of steps remaining before reaching a crosswalk's curb. Future research could delve into exploring how to best deliver such granular feedback to BLV users, alongside the necessary technological advancements needed to achieve this level of precision. These advantages, as our findings suggest, extend beyond merely improving navigation performance. Participants shared insights into how precise navigation could enhance their independence when navigating outdoors. It could empower BLV people to venture outdoors more frequently, unlocking new travel opportunities, as exemplified by P3's newfound confidence in using public transportation with StreetNav-like systems:

> "*I don't really use the city buses that much, except if I'm with somebody, but I tell you, [StreetNav] would make me want to get up, go outside, and walk to the bus stop.*" **–P3**

This newfound confidence is particularly noteworthy, considering the unpredictable nature of outdoor environments. Future research could explore new affordances that street camera-based systems can enable for people, in general.

3.6 Limitations

Our work revealed valuable insights into the benefits and effectiveness of a new approach that uses existing street cameras for outdoor navigation assistance. At the same time, we acknowledge that our work has several limitations.

StreetNav was developed using a single street camera and tested at a single street intersection. This approach means that there might be other technical hurdles and design considerations we didn't encounter due to the constraints of this setup. Future research could expand upon our design and investigate how street camera-based systems can adapt to different environments and challenges. Furthermore, to ensure the safety of participants and to fit the user study within a 120-minute timeframe, we designed the study routes to be less complex and dangerous. Real-world outdoor environments can vary significantly from one part of a city, state, or country to another. Our study location may not fully capture the diversity of scenarios BLV individuals encounter when navigating outdoors. Lastly, it's important to note that our study sample consisted of only eight BLV individuals. While their insights are valuable, their preferences for outdoor navigation may not represent the broader BLV community's perspectives. StreetNav was developed in response to the challenges identified in our formative study, but there could be additional challenges and design possibilities that we haven't explored. Future research should consider a more extensive and diverse participant pool to gain a more comprehensive understanding of the needs and preferences within the BLV community.

Conclusion

In this thesis, I leverage advancements in the field of computer vision to develop consolidated HCI applications that address a core element of the human experience: movement and our ability to perceive it. Through two interconnected research projects we help BLV people perceive sports gameplay from a third person perspective, and provide outdoor navigation assistance that helps BLV people interpret and interact with movement from a first-person perspective. Through these projects I demonstrate CV's capability to enhance human experiences related to movement.

These projects also highlight that enhancing people's capability to move and interpret movement amplifies their sense of agency and inspires within them new ideas for applications of such systems. Users envisioned scenarios in which Front Row, StreetNav, and extensions of these systems could improve their daily experiences, enhance their mobility, and increase their capacity for exploration of both physical and digital spaces.

In Chapter 2, we presented the Front Row system for automatically generating immersive audio representations of sports broadcast video, allowing BLV viewers to directly perceive what is happening in a tennis match rather than rely on others' descriptions. Our technical and user evaluations show Front Row's promise for making sports broadcasts accessible to BLV viewers, providing a more accurate understanding of gameplay and the agency to interpret the game themselves. Front Row's video-to-audio method can be integrated as a plug-in for video streaming platforms to make it possible for BLV people to access the vast repository of online sports and video content across the web.

In Chapter 3, we explored the idea of leveraging existing street cameras to support precise outdoor navigation for BLV pedestrians. Our resulting system, StreetNav, addresses BLV people's challenges in outdoor navigation when using GPS-based systems. Our user evaluation revealed StreetNav's potential to guide users to their destination more precisely than existing GPS-based systems. It also demonstrated its ability to offer real-time, context-aware navigation assistance, aiding in obstacle avoidance and safe street crossings. However, we also uncovered various considerations for street camera-based navigation systems, including challenges and opportunities in deploying such systems at scale. These challenges pave the way for future research to enhance the robustness and reliability of street camera-based navigation solutions. Our work highlights the untapped potential of embedding accessibility directly into urban infrastructure by leveraging existing resources, such as street cameras. We envision a future where such systems seamlessly integrate into urban environments, providing BLV individuals with safe and precise navigation capabilities, ultimately empowering them to confidently navigate their surroundings.

This thesis highlights the potential of CV-based HCI applications to enhance human movement experiences. Our findings also reveal an exciting prospect: extending similar principles and methodologies to other domains within artificial intelligence. Moving forward, I am keen on exploring the usage of language models and multi-modal approaches. Adopting this ethos, I am committed to developing innovative and inclusive AI solutions that center human experiences.

References

[1] J. B. Freeman, R. Dale, and T. A. Farmer, “Hand in motion reveals mind in motion,” *Frontiers in Psychology*, vol. 2, 2011.

[2] G. F. Koob, M. Le Moal, and R. F. Thompson, Eds., *Encyclopedia of behavioral neuroscience*, OCLC: 641281324, London: Academic Press, 2010, ISBN: 978-0-08-045396-5.

[3] N. Kuzik, P.-J. Naylor, J. C. Spence, and V. Carson, “Movement behaviours and physical, cognitive, and social-emotional development in preschool-aged children: Cross-sectional associations using compositional analyses,” *PLOS ONE*, vol. 15, no. 8, J. Brazo-Sayavera, Ed., e0237945, Aug. 18, 2020.

[4] M. Hofmann, J. Harris, S. E. Hudson, and J. Mankoff, “Helping hands: Requirements for a prototyping methodology for upper-limb prosthetics users,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA: ACM, May 7, 2016, pp. 1769–1780, ISBN: 978-1-4503-3362-7.

[5] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, “Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Florence Italy: ACM, Apr. 6, 2008, pp. 515–524, ISBN: 978-1-60558-011-1.

[6] R. Gouveia, F. Pereira, E. Karapanos, S. A. Munson, and M. Hassenzahl, “Exploring the design space of glanceable feedback for physical activity trackers,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg Germany: ACM, Sep. 12, 2016, pp. 144–155, ISBN: 978-1-4503-4461-6.

[7] L. M. Tang *et al.*, “Defining adherence: Making sense of physical activity tracker data,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–22, Mar. 26, 2018.

[8] M. Gillies, “What is movement interaction in virtual reality for?” In *Proceedings of the 3rd International Symposium on Movement and Computing*, Thessaloniki GA Greece: ACM, Jul. 5, 2016, pp. 1–4, ISBN: 978-1-4503-4307-7.

[9] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen, “Gaze + pinch interaction in virtual reality,” in *Proceedings of the 5th Symposium on Spatial User Interaction*, Brighton United Kingdom: ACM, Oct. 16, 2017, pp. 99–108, ISBN: 978-1-4503-5486-8.

[10] A. Radford *et al.*, *Learning Transferable Visual Models From Natural Language Supervision*, arXiv:2103.00020 [cs], Feb. 2021.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, Publisher: arXiv Version Number: 3.

[12] G. Jain *et al.*, "Front row: Automatically generating immersive audio representations of tennis broadcasts for blind viewers," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, San Francisco CA USA: ACM, Oct. 29, 2023, pp. 1–17, ISBN: 9798400701320.

[13] G. Jain *et al.*, "Towards accessible sports broadcasts for blind and low-vision viewers," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 19, 2023, pp. 1–7, ISBN: 978-1-4503-9422-2.

[14] G. Jain *et al.*, "Towards Accessible Sports Broadcasts for Blind and Low-Vision Viewers," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–7, ISBN: 978-1-4503-9422-2.

[15] G. Jain *et al.*, "StreetNav: Leveraging street cameras to support precise outdoor navigation for blind pedestrians," 2023, Publisher: arXiv Version Number: 1.

[16] B. Pettitt, K. Sharpe, and S. Cooper, "AUDETEL: Enhancing television for visually impaired people," *British Journal of Visual Impairment*, vol. 14, no. 2, pp. 48–52, May 1996, Publisher: SAGE Publications Ltd.

[17] S. Asakawa and A. Hurst, ""What just happened?": Understanding Non-visual Watching Sports Experiences," in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, Virtual Event USA: ACM, Oct. 2021, pp. 1–3, ISBN: 978-1-4503-8306-6.

[18] M. Cottril, *The Importance of Sports in Culture*, February 12, 2020.

[19] G. Jarvie, J. Thornton, and H. Mackie, *Sport, Culture and Society: An Introduction*, 3rd ed. Third edition. | Abingdon, Oxon ; New York, NY : Routledge is an imprint of the Taylor & Francis Group, an Informa Business, [2017]: Routledge, Jul. 2017.

[20] A. A. Raney and J. Bryant, *Handbook of Sports and Media*. Chapter 19: Why we watch and enjoy mediated sports., 2006.

[21] Action Audio, *Making Sports Broadcasts Accessible to People Living With Blindness or Low Vision*, 2021.

[22] C. Goncu and D. J. Finnegan, "'Did You See That!?' Enhancing the Experience of Sports Media Broadcast for Blind People," in *Human-Computer Interaction – INTERACT 2021*, vol. 12932, Cham: Springer International Publishing, 2021, pp. 396–417, ISBN: 978-3-030-85622-9 978-3-030-85623-6.

[23] H. Ohshima, M. Kobayashi, and S. Shimada, “Development of Blind Football Play-by-play System for Visually Impaired Spectators: Tangible Sports,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–6, ISBN: 978-1-4503-8095-9.

[24] E. Design, *Footbraile*, 2019.

[25] Santander, *Fieeld*, 2019.

[26] IrisVision, *IrisVision*, Retrieved July 15, 2022.

[27] H.-E. Innovations, 2001.

[28] T. Lin, “Hitting the Court, With an Ear on the Ball,” *The New York Times*, Jun. 2012.

[29] World Wide Web Consortium (W3C), *W3C Image Concepts*, 2022.

[30] T. McEwan and B. Weerts, “ALT Text and Basic Accessibility,” Sep. 2007.

[31] World Wide Web Consortium (W3C), *Making Audio and Video Media Accessible*, 2022.

[32] American Council of the Blind, *The Audio Description Project*, 2022.

[33] K. Mack, E. Cutrell, B. Lee, and M. R. Morris, “Designing Tools for High-Quality Alt Text Authoring,” in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS ’21, New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 1–14, ISBN: 978-1-4503-8306-6.

[34] C. Gleason, A. Pavel, H. Gururaj, K. Kitani, and J. Bigham, “Making GIFs Accessible,” in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS ’20, New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 1–10, ISBN: 978-1-4503-7103-2.

[35] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To, ““It’s Complicated”: Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–19, ISBN: 978-1-4503-8096-6.

[36] A. Stangl, M. R. Morris, and D. Gurari, “"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20, New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–13, ISBN: 978-1-4503-6708-0.

[37] X. Liu, P. Carrington, X. A. Chen, and A. Pavel, “What Makes Videos Accessible to Blind and Visually Impaired People?” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–14, ISBN: 978-1-4503-8096-6.

[38] Y. Wang, R. Wang, C. Jung, and Y.-S. Kim, “What makes web data tables accessible? Insights and a tool for rendering accessible tables for people with visual impairments,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–20, ISBN: 978-1-4503-9157-3.

[39] K. Angerbauer *et al.*, “Accessibility for Color Vision Deficiencies: Challenges and Findings of a Large Scale Study on Paper Figures,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–23, ISBN: 978-1-4503-9157-3.

[40] Y. Wang, W. Liang, H. Huang, Y. Zhang, D. Li, and L.-F. Yu, “Toward Automatic Audio Description Generation for Accessible Videos,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–12, ISBN: 978-1-4503-8096-6.

[41] R.-C. Chang *et al.*, “OmniScribe: Authoring Immersive Audio Descriptions for 360 ° Videos,” p. 14, 2022.

[42] C. Gleason *et al.*, “Twitter A11y: A Browser Extension to Make Twitter Images Accessible,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–12, ISBN: 978-1-4503-6708-0.

[43] M. R. Zhang, M. Zhong, and J. O. Wobbrock, “Ga11y: An Automated GIF Annotation System for Visually Impaired Users,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–16, ISBN: 978-1-4503-9157-3.

[44] A. Sharif, O. H. Wang, A. T. Muongchan, K. Reinecke, and J. O. Wobbrock, “VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–19, ISBN: 978-1-4503-9157-3.

[45] M. R. Morris, J. Johnson, C. L. Bennett, and E. Cutrell, “Rich Representations of Visual Content for Screen Reader Users,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–11, ISBN: 978-1-4503-5620-6.

[46] V. Potluri, T. E. Grindeland, J. E. Froehlich, and J. Mankoff, “Examining Visual Semantic Understanding in Blind and Low-Vision Technology Users,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–14, ISBN: 978-1-4503-8096-6.

[47] K. Rector, K. Salmon, D. Thornton, N. Joshi, and M. R. Morris, “Eyes-Free Art: Exploring Proxemic Audio Interfaces For Blind and Low Vision Art Engagement,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, Sep. 2017.

[48] J. Snyder, “Audio description: The visual made verbal,” *International Congress Series*, vol. 1282, pp. 935–939, Sep. 2005.

[49] S. Asakawa *et al.*, “An Independent and Interactive Museum Experience for Blind People,” in *Proceedings of the 16th International Web for All Conference*, San Francisco CA USA: ACM, May 2019, pp. 1–9, ISBN: 978-1-4503-6716-5.

[50] J. Shin, J. Cho, and S. Lee, “Please Touch Color: Tactile-Color Texture Design for The Visually Impaired,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–7, ISBN: 978-1-4503-6819-3.

[51] F. M. Li *et al.*, *Understanding Visual Arts Experiences of Blind People*, arXiv:2301.12687 [cs], Jan. 2023.

[52] V. Nair *et al.*, “NavStick: Making Video Games Blind-Accessible via the Ability to Look Around,” in *The 34th Annual ACM Symposium on User Interface Software and Technology*, Virtual Event USA: ACM, Oct. 2021, pp. 538–551, ISBN: 978-1-4503-8635-7.

[53] B. A. Smith and S. K. Nayar, “The RAD: Making Racing Games Equivalently Accessible to People Who Are Blind,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’18, New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–12, ISBN: 978-1-4503-5620-6.

[54] K. Mack, D. Bragg, M. R. Morris, M. W. Bos, I. Albi, and A. Monroy-Hernández, “Social App Accessibility for Deaf Signers,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–31, Oct. 2020.

[55] N. Staggers and D. Kobus, “Comparing Response Time, Errors, and Satisfaction Between Text-based and Graphical User Interfaces During Nursing Order Tasks,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 7, no. 2, pp. 164–176, 2000.

[56] M. Butler, L. M. Holloway, S. Reinders, C. Goncu, and K. Marriott, “Technology developments in touch-based accessible graphics: A systematic review of research 2010-2020,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21, New York, NY, USA: Association for Computing Machinery, 2021, ISBN: 978-1-4503-8096-6.

[57] A. Reichinger, S. Maierhofer, and W. Purgathofer, “High-quality tactile paintings,” *Journal on Computing and Cultural Heritage*, vol. 4, no. 2, pp. 1–13, Nov. 2011.

[58] C. Goncu, A. Madugalla, S. Marinai, and K. Marriott, “Accessible On-Line Floor Plans,” in *Proceedings of the 24th International Conference on World Wide Web*, Florence Italy: International World Wide Web Conferences Steering Committee, May 2015, pp. 388–398, ISBN: 978-1-4503-3469-3.

[59] A. Madugalla, K. Marriott, S. Marinai, S. Capobianco, and C. Goncu, “Creating Accessible Online Floor Plans for Visually Impaired Readers,” *ACM Transactions on Accessible Computing*, vol. 13, no. 4, pp. 1–37, Oct. 2020.

[60] D. Prescher, J. Bornschein, W. Kohlmann, and G. Weber, “Touching graphical applications: Bimanual tactile interaction on the HyperBraille pin-matrix display,” *Universal Access in the Information Society*, vol. 17, no. 2, pp. 391–409, Jun. 2018.

[61] C. Goncu and K. Marriott, “GraVVITAS: Generic Multi-touch Presentation of Accessible Graphics,” in *Human-Computer Interaction – INTERACT 2011*, D. Hutchison *et al.*, Eds., vol. 6946, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 30–48, ISBN: 978-3-642-23773-7 978-3-642-23774-4.

[62] S. K. Kane, M. R. Morris, and J. O. Wobbrock, “Touchplates: Low-cost tactile overlays for visually impaired touch screen users,” in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, Bellevue Washington: ACM, Oct. 2013, pp. 1–8, ISBN: 978-1-4503-2405-2.

[63] H. Iwata, H. Yano, F. Nakaizumi, and R. Kawamura, “Project FEELEX: Adding haptic surface to graphics,” in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH ’01*, Not Known: ACM Press, 2001, pp. 469–476, ISBN: 978-1-58113-374-5.

[64] J. Lee, J. Herskovitz, Y.-H. Peng, and A. Guo, “ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–15, ISBN: 978-1-4503-9157-3.

[65] A. F. Siu, S. Kim, J. A. Miele, and S. Follmer, “shapeCAD: An Accessible 3D Modelling Workflow for the Blind and Visually-Impaired Via 2.5D Shape Displays,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh PA USA: ACM, Oct. 2019, pp. 342–354, ISBN: 978-1-4503-6676-2.

[66] O. Falase, A. F. Siu, and S. Follmer, “Tactile Code Skimmer: A Tool to Help Blind Programmers Feel the Structure of Code,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh PA USA: ACM, Oct. 2019, pp. 536–538, ISBN: 978-1-4503-6676-2.

[67] R. Shilkrot, J. Huber, C. Liu, P. Maes, and S. C. Nanayakkara, “FingerReader: A wearable device to support text reading on the go,” in *CHI ’14 Extended Abstracts on Human Factors*

in Computing Systems, Toronto Ontario Canada: ACM, Apr. 2014, pp. 2359–2364, ISBN: 978-1-4503-2474-8.

[68] R. Shilkrot, J. Huber, W. Meng Ee, P. Maes, and S. C. Nanayakkara, “FingerReader: A Wearable Device to Explore Printed Text on the Go,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea: ACM, Apr. 2015, pp. 2363–2372, ISBN: 978-1-4503-3145-6.

[69] L. Stearns, V. DeSouza, J. Yin, L. Findlater, and J. E. Froehlich, “Augmented Reality Magnification for Low Vision Users with the Microsoft Hololens and a Finger-Worn Camera,” in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, Baltimore Maryland USA: ACM, Oct. 2017, pp. 361–362, ISBN: 978-1-4503-4926-0.

[70] G. Hamilton-Fletcher, M. Obrist, P. Watten, M. Mengucci, and J. Ward, “"I Always Wanted to See the Night Sky": Blind User Preferences for Sensory Substitution Devices,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA: ACM, May 2016, pp. 2162–2174, ISBN: 978-1-4503-3362-7.

[71] Peter Meijer, *The vOICe*, Retrieved August 2022.

[72] A. Siu, G. S-H Kim, S. O’Modhrain, and S. Follmer, “Supporting Accessible Data Visualization Through Audio Data Narratives,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–19, ISBN: 978-1-4503-9157-3.

[73] L. M. Holloway, C. Goncu, A. Ilsar, M. Butler, and K. Marriott, “Infosonics: Accessible Infographics for People who are Blind using Sonification and Voice,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–13, ISBN: 978-1-4503-9157-3.

[74] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, “Football Action Recognition Using Hierarchical LSTM,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 155–163, ISBN: 978-1-5386-0733-6.

[75] Y.-C. Huang, I.-N. Liao, C.-H. Chen, T.-U. İk, and W.-C. Peng, “TrackNet: A Deep Learning Network for Tracking High-speed and Tiny Objects in Sports Applications*,” in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, ISSN: 2643-6213, Sep. 2019, pp. 1–8.

[76] Y. Nishikawa, H. Sato, and J. Ozawa, “Multiple sports player tracking system based on graph optimization using low-cost cameras,” in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, ISSN: 2158-4001, Jan. 2018, pp. 1–4.

[77] Z. Chen *et al.*, “Augmenting Sports Videos with VisCommentator,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 824–834, Jan. 2022.

[78] A. Ghosh and C. V. Jawahar, “SmartTennisTV: Automatic indexing of tennis videos,” *arXiv:1801.01430 [cs]*, Jan. 2018, arXiv: 1801.01430.

[79] A. Ghosh, S. Singh, and C. V. Jawahar, “Towards Structured Analysis of Broadcast Badminton Videos,” *arXiv:1712.08714 [cs]*, Dec. 2017, arXiv: 1712.08714.

[80] R. Voeikov, N. Falaleev, and R. Baikulov, “TTNet: Real-time temporal and spatial video analysis of table tennis,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 3866–3874, ISBN: 978-1-72819-360-1.

[81] J. C. Flanagan, “The critical incident technique,” *Psychological Bulletin*, vol. 51, no. 4, pp. 327–358, 1954.

[82] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006.

[83] NVivo, *NVivo*, 1997.

[84] Valve Corporation, *Steam Audio*, 2018.

[85] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. J. Ross Publishing, Jul. 2013, Google-Books-ID: fvDLCgAAQBAJ, ISBN: 978-1-60427-070-9.

[86] J. Matas, C. Galambos, and J. Kittler, “Progressive Probabilistic Hough Transform,” in *Procedings of the British Machine Vision Conference 1998*, Southampton: British Machine Vision Association, 1998, pp. 26.1–26.10, ISBN: 978-1-901725-04-9.

[87] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” 2015, pp. 3431–3440.

[88] G. Jocher et al., *Yolov5*, April 2021.

[89] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997, Publisher: MIT press.

[90] A. Graves, *Generating Sequences With Recurrent Neural Networks*, arXiv:1308.0850 [cs], Jun. 2014.

[91] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[92] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, San Diego, CA, USA: IEEE, 2005, pp. 886–893, ISBN: 978-0-7695-2372-9.

[93] K. A. Hunt, T. Bristol, and R. E. Bashaw, "A conceptual approach to classifying sports fans," *The Journal of Services Marketing*, vol. 13, no. 6, pp. 439–452, 1999.

[94] L. A. Goodman, "Snowball Sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, 1961.

[95] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, ser. Human Mental Workload, P. A. Hancock and N. Meshkati, Eds., vol. 52, North-Holland, Jan. 1988, pp. 139–183.

[96] J. Garhammer and H. Newton, "Applied Video Analysis for Coaches: Weightlifting Examples," *International Journal of Sports Science & Coaching*, vol. 8, no. 3, pp. 581–594, Sep. 2013, Publisher: SAGE Publications.

[97] MIPsoft, *BlindSquare*, 2016.

[98] M. Inc., *Microsoft Soundscape - Microsoft Research.* `https://www.microsoft.com/en-us/research/product/soundscape/`. (2018), 2018.

[99] Sendero Group, *Seeing Eye GPS*, 2019.

[100] The Royal Institution for the Advancement of Learning (McGill University), *Autour*, 2017.

[101] H. Kacorri *et al.*, "Insights on Assistive Orientation and Mobility of People with Visual Impairment Based on Large-Scale Longitudinal Data," *ACM Transactions on Accessible Computing*, vol. 11, no. 1, pp. 1–28, Apr. 2018.

[102] M. Saha, A. J. Fiannaca, M. Kneisel, E. Cutrell, and M. R. Morris, "Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh PA USA: ACM, Oct. 2019, pp. 222–235, ISBN: 978-1-4503-6676-2.

[103] GPS.gov, *GPS Accuracy*, 2022.

[104] M. Modsching, R. Kramer, and K. ten Hagen, "Field trial on gps accuracy in a medium size city: The influence of built-up," in *3rd workshop on positioning, navigation and communication*, vol. 2006, 2006, pp. 209–218.

[105] F. Van Diggelen and P. Enge, “The world’s first GPS MOOC and worldwide laboratory using smartphones,” vol. 1, 2015, pp. 361–369, ISBN: 978-1-5108-1725-8.

[106] S. A. Paneels, D. Varenne, J. R. Blum, and J. R. Cooperstock, “The Walking Straight Mobile Application: Helping the Visually Impaired Avoid Veering,” Jul. 2013.

[107] J. Pariti, V. Tibdewal, and T. Oh, “Intelligent Mobility Cane - Lessons Learned from Evaluation of Obstacle Notification System using a Haptic Approach,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–8, ISBN: 978-1-4503-6819-3.

[108] G. Presti *et al.*, “WatchOut: Obstacle Sonification for People with Visual Impairment or Blindness,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh PA USA: ACM, Oct. 2019, pp. 402–413, ISBN: 978-1-4503-6676-2.

[109] M. Avila and L. Zeng, “A Survey of Outdoor Travel for Visually Impaired People Who Live in Latin-American Region,” in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA ’17, New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 9–12, ISBN: 978-1-4503-5227-7.

[110] D. Ahmetovic, C. Gleason, C. Ruan, K. Kitani, H. Takagi, and C. Asakawa, “NavCog: A navigational cognitive assistant for the blind,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ser. MobileHCI ’16, New York, NY, USA: Association for Computing Machinery, Sep. 2016, pp. 90–99, ISBN: 978-1-4503-4408-1.

[111] M. Nakajima and S. Haruyama, “Indoor navigation system for visually impaired people using visible light communication and compensated geomagnetic sensing,” in *2012 1st IEEE International Conference on Communications in China (ICCC)*, 2012, pp. 524–529.

[112] T. Gallagher, E. Wise, B. Li, A. G. Dempster, C. Rizos, and E. Ramsey-Stewart, “Indoor positioning system based on sensor fusion for the blind and visually impaired,” in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2012, pp. 1–9.

[113] J.-E. Kim, M. Bessho, S. Kobayashi, N. Koshizuka, and K. Sakamura, “Navigating visually impaired travelers in a large train station using smartphone and bluetooth low energy,” in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, ser. SAC ’16, New York, NY, USA: Association for Computing Machinery, Apr. 2016, pp. 604–611, ISBN: 978-1-4503-3739-7.

[114] D. Sato *et al.*, “NavCog3 in the Wild: Large-scale Blind Indoor Navigation Assistant with Semantic Features,” *ACM Transactions on Accessible Computing*, vol. 12, no. 3, pp. 1–30, Sep. 2019.

[115] Amnesty International, *Surveillance City: NYPD can use more than 15,000 cameras to track people using facial recognition in Manhattan, Bronx and Brooklyn*, 2021.

[116] Laura Griffin, *Surveillance Cameras Are Everywhere. And They’re Only Going To Get More Ubiquitous*, 2020.

[117] Frank Hersey, *China to have 626 million surveillance cameras within 3 years*, 2017.

[118] Coco Feng, *China the most surveilled nation? The US has the largest number of CCTV cameras per capita*, 2019.

[119] Liza Lin, Newley Purnell, *A World With a Billion Cameras Watching You Is Just Around the Corner*, 2019.

[120] C. Yoon *et al.*, “Leveraging Augmented Reality to Create Apps for People with Visual Disabilities: A Case Study in Indoor Navigation,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh PA USA: ACM, Oct. 2019, pp. 210–221, ISBN: 978-1-4503-6676-2.

[121] C. Vicek, P. McLain, and M. Murphy, “Gps/dead reckoning for vehicle tracking in the" urban canyon" environment,” in *Proceedings of VNIS’93-Vehicle Navigation and Information Systems Conference*, IEEE, 1993, pp. 461–34.

[122] D. L. Harkey, D. L. Carter, J. M. Barlow, and B. L. Bentzen, “Accessible pedestrian signals: A guide to best practices,” *National Cooperative Highway Research Program, Contractor’s Guide for NCHRP Project*, 2007.

[123] H. Son, D. Krishnagiri, V. S. Jeganathan, and J. Weiland, “Crosswalk Guidance System for the Blind,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, ISSN: 2694-0604, Jul. 2020, pp. 3327–3330.

[124] X. Li, H. Cui, J.-R. Rizzo, E. Wong, and Y. Fang, “Cross-Safe: A Computer Vision-Based Approach to Make All Intersection-Related Pedestrian Signals Accessible for the Visually Impaired,” in *Advances in Computer Vision*, K. Arai and S. Kapoor, Eds., vol. 944, Cham: Springer International Publishing, 2020, pp. 132–146, ISBN: 978-3-030-17797-3 978-3-030-17798-0.

[125] R. Guy and K. Truong, “CrossingGuard: Exploring information content in navigation aids for visually impaired pedestrians,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin Texas USA: ACM, May 2012, pp. 405–414, ISBN: 978-1-4503-1015-4.

[126] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, Singapore: IEEE, May 2017, pp. 6533–6540, ISBN: 978-1-5090-4633-1.

[127] P.-J. Duh, Y.-C. Sung, L.-Y. F. Chiang, Y.-J. Chang, and K.-W. Chen, "V-eye: A vision-based navigation system for the visually impaired," *IEEE Transactions on Multimedia*, vol. 23, pp. 1567–1580, 2020.

[128] Y. Lin, K. Wang, W. Yi, and S. Lian, "Deep Learning Based Wearable Assistive System for Visually Impaired People," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 2549–2557, ISBN: 978-1-72815-023-9.

[129] L. Ran, S. Helal, and S. Moore, "Drishti: An integrated indoor/outdoor blind navigation system and service," in *Second IEEE Annual Conference on Pervasive Computing and Communications, 2004. Proceedings of the*, IEEE, 2004, pp. 23–30.

[130] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, 2018.

[131] A. Fiannaca, I. Apostolopoulous, and E. Folmer, "Headlock: A wearable navigation aid that helps blind cane users traverse large open spaces," in *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, ser. ASSETS '14, New York, NY, USA: Association for Computing Machinery, Oct. 2014, pp. 323–324, ISBN: 978-1-4503-2720-6.

[132] W.-C. Chang, C.-H. Wu, W.-T. Luo, and H.-C. Ling, "Mobile robot navigation and control with monocular surveillance cameras," in *2013 CACS International Automatic Control Conference (CACS)*, Dec. 2013, pp. 192–197.

[133] P. Oščádal, D. Huczala, J. Bém, V. Krys, and Z. Bobovský, "Smart Building Surveillance System as Shared Sensory System for Localization of AGVs," *Applied Sciences*, vol. 10, no. 23, p. 8452, Nov. 2020.

[134] R. Pflugfelder and H. Bischof, "Localization and Trajectory Reconstruction in Surveillance Cameras with Nonoverlapping Views," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 709–21, Apr. 2010.

[135] J. Shim and Y. Cho, "A Mobile Robot Localization via Indoor Fixed Remote Surveillance Cameras," *Sensors*, vol. 16, no. 2, p. 195, Feb. 2016.

[136] J.-H. Shim and Y.-I. Cho, "A mobile robot localization using external surveillance cameras at indoor," *Procedia Computer Science*, vol. 56, pp. 502–507, 2015.

[137] S. Mascetti, D. Ahmetovic, A. Gerino, and C. Bernareggi, “ZebraRecognizer: Pedestrian crossing recognition for people with visual impairment or blindness,” *Pattern Recognition*, vol. 60, pp. 405–419, Dec. 2016.

[138] D. Ahmetovic, R. Manduchi, J. M. Coughlan, and S. Mascetti, “Mind Your Crossings: Mining GIS Imagery for Crosswalk Localization,” *ACM Transactions on Accessible Computing*, vol. 9, no. 4, pp. 1–25, Dec. 2017.

[139] MQTT, *MQTT: The Standard for IoT Messaging*, 2022.

[140] Apple Inc., *Swift*, 2023.

[141] Apple Inc., *VoiceOver*, 2023.

[142] Apple Inc., *Apple Maps*, 2023.

[143] Nvidia, *Nvidia DeepStream GStreamer Plugin: NvDCF Tracker*, 2023.

[144] J. Terven and D. Cordova-Esparza, “A comprehensive review of yolo: From yolov1 to yolov8 and beyond,” *arXiv preprint arXiv:2304.00501*, 2023.

[145] NVIDIA, *NVIDIA deepstream sdk developer guide*, 2023.

[146] W. Taymans, S. Baker, A. Wingo, R. S. Bultje, and S. Kost, “Gstreamer application development manual (1.2. 3),” *Publicado en la Web*, vol. 72, 2013.

[147] M. Contributors, *Openmmlab’s next generation video understanding toolbox and benchmark*, `https://github.com/open-mmlab/mmaction2`, 2020.

[148] J. Guerreiro, D. Sato, S. Asakawa, H. Dong, K. M. Kitani, and C. Asakawa, “CaBot: Designing and Evaluating an Autonomous Navigation Robot for Blind People,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS ’19, New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 68–82, ISBN: 978-1-4503-6676-2.

[149] F. Duchoň *et al.*, “Path Planning with Modified a Star Algorithm for a Mobile Robot,” *Procedia Engineering*, vol. 96, pp. 59–69, 2014.

[150] S. Kayukawa *et al.*, “BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*, Glasgow, Scotland Uk: ACM Press, 2019, pp. 1–12, ISBN: 978-1-4503-5970-2.

[151] J. Guerreiro, D. Sato, D. Ahmetovic, E. Ohn-Bar, K. M. Kitani, and C. Asakawa, “Virtual navigation for blind people: Transferring route knowledge to the real-World,” *International Journal of Human-Computer Studies*, vol. 135, p. 102 369, Mar. 2020.

[152] G. Jain, Y. Teng, D. H. Cho, Y. Xing, M. Aziz, and B. A. Smith, “"I Want to Figure Things Out": Supporting Exploration in Navigation for People with Visual Impairments,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, 63:1–63:28, Apr. 2023.

[153] N. N. Abd Hamid and A. D. Edwards, “Facilitating route learning using interactive audio-tactile maps for blind and visually impaired people,” in *CHI ’13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA ’13*, Paris, France: ACM Press, 2013, p. 37, ISBN: 978-1-4503-1952-2.