Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts for Blind Viewers

Gaurav Jain
Columbia University
New York, NY, USA
gaurav@cs.columbia.edu

Basel Hindi
Columbia University
New York, NY, USA
basel.hindi@columbia.edu

Connor Courtien*
Hunter College
New York, NY, USA
cjcourtien@gmail.com

Xin Yi Therese Xu*
Pomona College
Claremont, CA, USA
xx2449@columbia.edu

Conrad Wyrick*
University of Florida
Gainesville, FL, USA
conradwyrick@ufl.edu

Michael Malcolm*
SUNY at Albany
Albany, NY, USA
mmalcolm@albany.edu

Brian A. Smith
Columbia University
New York, NY, USA
brian@cs.columbia.edu

Figure 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.

ABSTRACT

Blind and low-vision (BLV) people face challenges watching sports due to the lack of accessibility of sports broadcasts. Currently, BLV people rely on descriptions from TV commentators, radio announcers, or their friends to understand the game. These descriptions, however, do not allow BLV viewers to visualize the action by themselves. We present *Front Row*, a system that automatically generates an immersive audio representation of sports broadcasts, specifically tennis, allowing BLV viewers to more directly perceive what is happening in the game. Front Row first recognizes gameplay from the video feed using computer vision, 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' moods and strategies. We discuss future implications of Front Row and illustrate several applications, including a Front Row plug-in for video streaming platforms to enable BLV people to visualize the action in sports videos across the Web.

*This work was done while Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, and Michael Malcolm were interns at Columbia University.

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606830

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; *Accessibility technologies.*

KEYWORDS

Visual impairments, sports, accessibility, computer vision

ACM Reference Format:

Gaurav Jain, Basel Hindi, Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, Michael Malcolm, and Brian A. Smith. 2023. Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts for Blind Viewers. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3586183.3606830

1 INTRODUCTION

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 [5, 56], making the experience of watching sports exclusionary and isolating for them [11, 36, 59]. 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 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 2a) convey players' positions and actions thoroughly, allowing viewers to focus on the parts of the game they find interesting. The radio descriptions (Figure 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 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 "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 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 [1, 13, 21, 33, 53, 62]. For example, Action Audio [1] acquires the ball's position using a specialized tracking system [32] 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. 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 8, 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 [41] 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 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 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") [48, 80] and videos via audio descriptions (AD) [2, 79]. Many researchers have studied ways to create effective descriptions for BLV people, by proposing methods and guidelines for authoring descriptions [3, 6, 19, 42, 45, 72, 78] as well as by introducing tools to support and automate the process [9, 20, 63, 77, 78, 82]. Prior work, however, shows that descriptions do not provide BLV people a spatial understanding of the visual content [49, 57, 60]. Spatial

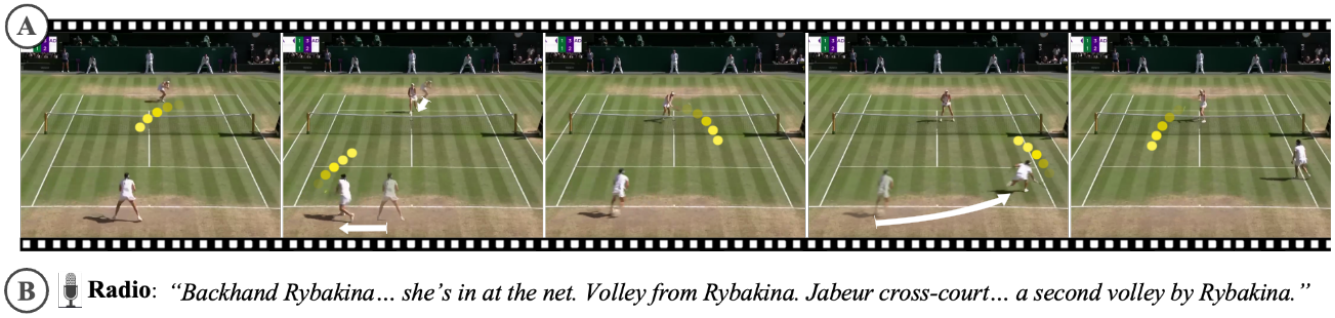


Figure 2: Tennis gameplay as experienced (a) on TV via visuals by sighted viewers, and (b) on radio via announcer's description by BLV viewers. The visuals allow sighted viewers to perceive players' positions and actions to fully understand gameplay. The ball's path is indicated in yellow, and the players' 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.

understanding of the visual content is crucial for interacting with rich visual media, such as for watching TV [2, 56, 70], exploring museums [4, 40, 60, 66], playing video games [50, 69], and engaging with social media [20, 44, 49, 82].

Another approach to visual media accessibility is using tactile graphics, which conveys spatial information via touch [8, 71]. Tactile graphics have been successfully used to understand the spatial layout of paintings [61], floor plans [22, 46], and more [14, 23, 34, 38, 39, 58, 68]. Prior work has also explored finger-worn devices [64, 65, 73] 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 [26, 55], as well as for particular forms of images such as time series charts [29, 67]. However, limited work has been done to make videos [9], 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.

2.2 Sports Broadcast Accessibility

Sports play an important role in enhancing people's social and cultural lives [11, 36, 59]. However, BLV people often experience sports in isolation because many existing sports broadcasts remain inaccessible to them [5]. Past research has explored different ways of making sports broadcasts accessible to BLV people, leveraging tactile graphic displays for football games [13, 53, 62] and 3D spatialized audio for tennis games [1, 21].

Most approaches, however, rely on specialized hardware which may not always be feasible. For example, Action Audio [1] requires the court to be equipped with the Hawk-Eye ball tracking technology [32], 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 [35], 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.

2.3 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 [30], and players [51] and developing new applications using them [10, 17, 18, 76]. For instance, Voeikov et al. [76] introduced a deep learning-based system for automatic refereeing in table tennis games. Ghosh et al. [18] 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.

3 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?

3.1 Methods

Participants. We recruited five BLV participants (three males, two females; aged 23–60) by posting to social media platforms. Table 1 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) [15], 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 [7] involving two members of our research team. Each researcher independently reviewed the interview transcripts to generate an initial set of codes using NVivo [52]. 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.

3.2 Understanding BLV Viewers' Challenges of Watching Sports

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

3.2.1 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 [11, 36, 59] and that the sense of shared excitement and affiliation is a big motivation for BLV people to watch sports [5].

3.2.2 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.

3.3 Understanding BLV Viewers' Information Preferences for Watching Sports

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

3.3.1 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).

3.3.2 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).

3.4 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 3.3.1), 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 3.2.2), 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 3.2.1), 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 3.3.2, 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.

4 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**).

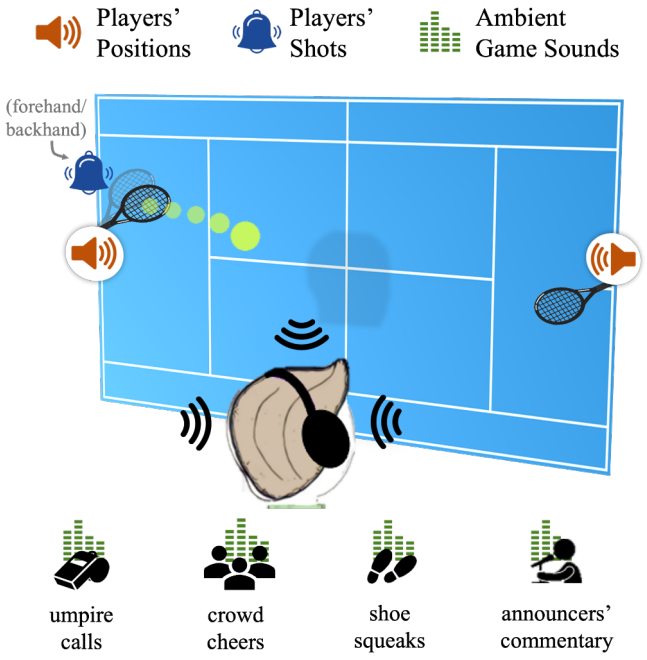


Figure 3: Front Row's 3D soundscape. The tennis court is displayed on a 2D plane orthogonal to the BLV viewer. Players' positions are represented by continuous humming sounds, and players' shots are represented by bell sounds similar to those in blind tennis [41]. These sounds are blended with the TV broadcast's original audio to incorporate ambient noises and the announcers' commentary.

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 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 [75], which provides a built-in head-related transfer function (HRTF) [81]. Our design for Front Row resulted from several co-design sessions with our two BLV co-authors and consideration of prior work [1, 21]. In the following subsections, we describe each of Front Row's three sound cues.

4.1 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 [50, 69]. 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 [1], 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.

4.2 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 they hit the shot. We chose a bell sound because it resembles the sound of the ball used in blind tennis [41], which was also the choice in prior work [1, 21]. 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.

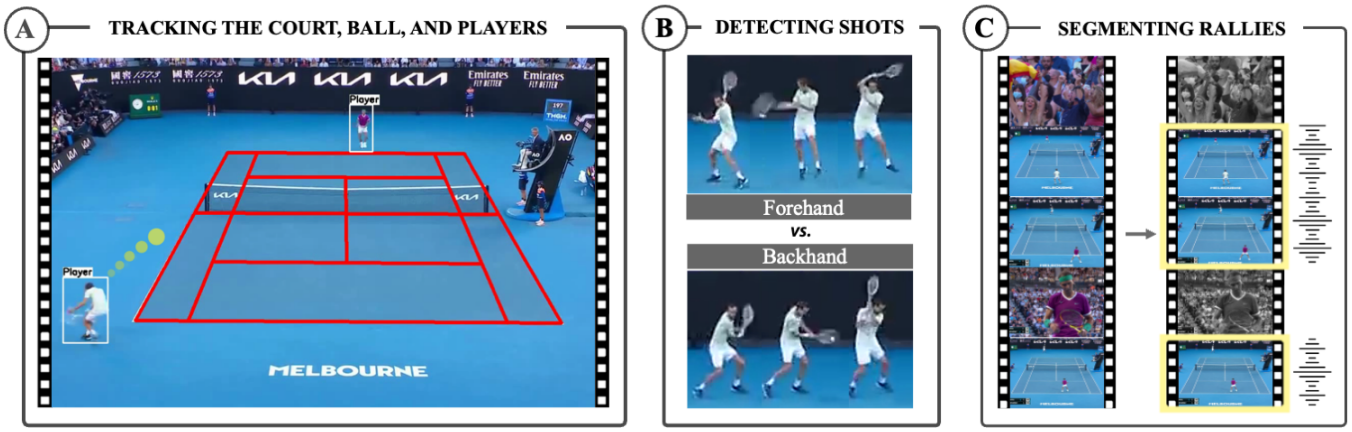


Figure 4: Front Row's computer vision pipeline. The pipeline takes as input only the tennis broadcast video feed to generate audio representations of the game. It consists of three components: (a) tracking the court, ball, and players; (b) detecting shots: recognizing when, where, and how players hit 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.

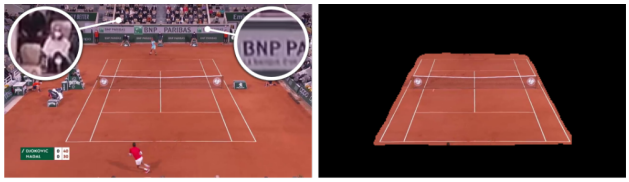
4.3 Blending Ambient Game Sounds

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**.

5 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 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.



(a) (b)

Figure 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.

5.1 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 [47] 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.

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 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 [43], masking out the background as Figure 5b shows. We then detect white lines in the roughly masked court area only. This fix eliminated false detections of white lines completely.

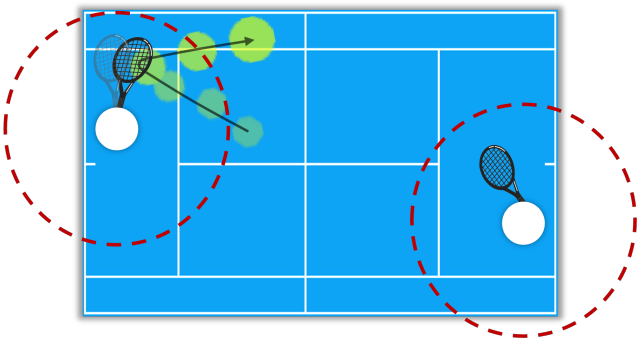


Figure 6: Illustration of technique for detecting *when* a shot is hit. Change in the direction of the ball's trajectory within a fixed radial distance from either player is used to identify *when* shots are hit by a player. This change in direction is computed using player and ball coordinates with respect to the court's 2D representation.

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 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 [30]. 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 5.2, 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 [37] to find the players' positions in terms of pixel coordinates. We chose YOLOv5 [37] 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.

5.2 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 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 [25, 28]). 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.

5.3 Segmenting Rallies

With the components from Section 5.1 and Section 5.2, 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 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) [54] that detects the full view of the court in broadcast videos. The SVC takes as input the histogram of oriented gradients (HOG) [12] 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 4C. The ambient game sounds, however, can be heard at all times during the game, even when the ball is not in play.

6 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?*

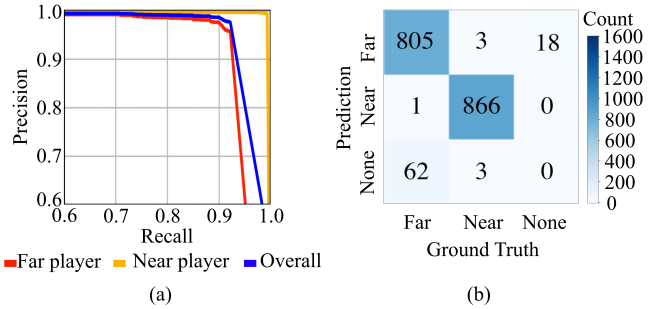


Figure 7: Player tracking accuracy results. (a) The precision-recall curve at 0.5 IoU threshold. (b) The Confusion matrix at 0.5 IoU threshold and 0.5 confidence threshold. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model achieves a 97.2% mean average precision at 0.5 IoU threshold.

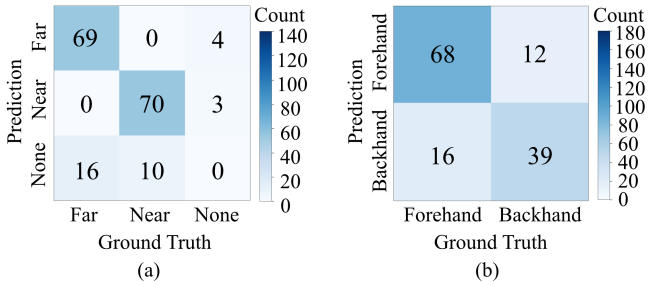


Figure 8: Confusion matrices for (a) detecting occurrence of shots and for (b) detecting the type of shots, i.e., forehands vs. backhands. Far and Near refer to the two players, with Far referring to the player farther away in the TV broadcast's camera view. Our model correctly detects the occurrence of 80.8% shots, and classifies the shot types with 79.3% accuracy.

6.1 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 1, (ii) grass: the green court in Figure 2, and (iii) clay: the red court in Figure 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 7. 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.

6.2 Results

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

Player Tracking Accuracy. Figure 7 summarizes the accuracy with which Front Row tracks players' positions via a precision-recall curve (Figure 7a) and confusion matrix (Figure 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 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 [30]. Future improvements in ball tracking could potentially increase shot detection accuracy.

Shot Type Detection Accuracy. Figure 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.

Table 1: Self-reported demographics of our study participants. Five BLV participants (F1–F5) were recruited for the formative study (Section 3), while twelve BLV participants (P1–P12) were part of the user study evaluating Front Row (Section 7). 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 [31] 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

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 4.3) 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 4.1) 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.

7 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 7.2). Second, we wanted to quantitatively analyse BLV people's overall experience of viewing tennis games using Front Row and these existing means (Section 7.3). 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 7.4).

7.1 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 [24]. 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 [31], and familiarity with tennis rules.

Table 1 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 [27]. We chose the objective measures for gameplay understanding and the subjective measure of participants' overall experience based on our formative study findings (Section 3).

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.

7.2 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 9 shows participants' available choices for these two questions and how we scored participants' responses. As Figure 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 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.

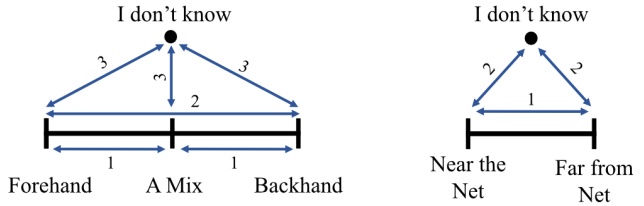
We computed participants' error rates by calculating the distance of their response from the correct answer on the relevant spectrum from Figure 9. Note that we penalized 'I don't know' more strongly — with a greater distance value — since it reflected them not being able to ascertain any information at all.

Recognizing Players' Predominant Shot Types. Figure 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:

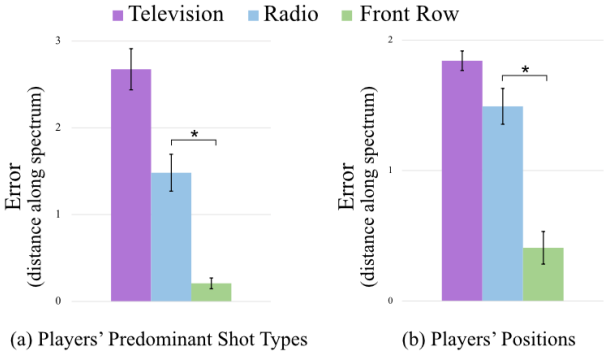
> P3: *"I understand that there's a couple backhands and then there's a forehand, but I don't know who's doing what. ... [The announcer] was mostly talking about like every third shot, so it's confusing."*

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



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

Figure 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.



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

Figure 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.

> 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."*

Identifying Players' Positions. Figure 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."*

7.3 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 3).

Perceived Information Overload. Figure 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 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

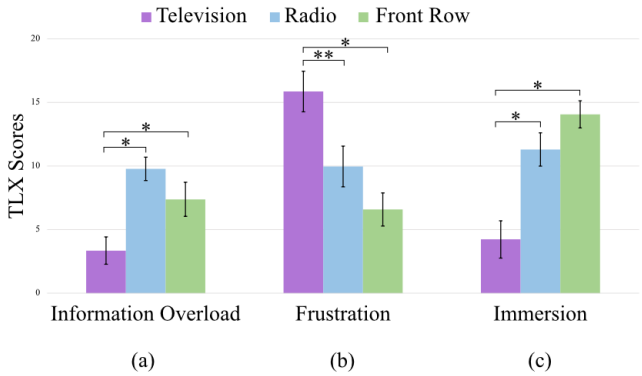


Figure 11: Average TLX scores for participants' overall experience. Participants rated their (a) perceived information overload, (b) perceived frustration, and (c) perceived immersion on a scale of 1—20 while viewing tennis rallies via the three audio formats. The error bars indicate standard error. 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.

($p < 0.01$). However, there was no significant difference between Radio and Front Row in the post-hoc analysis.

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 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."*

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."*

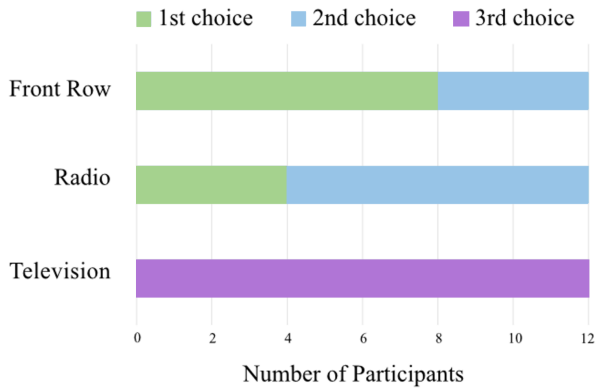
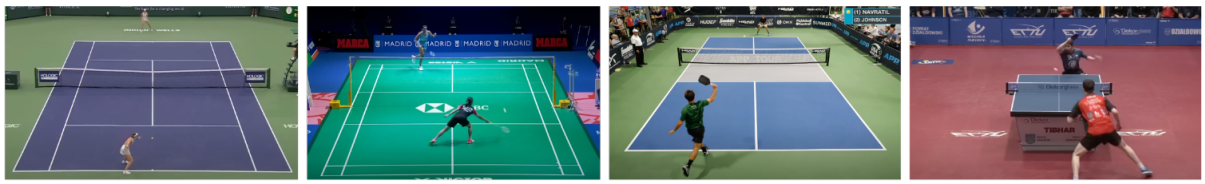


Figure 12: Forced ranking results. Participants ranked the three audio formats in order of their preference for viewing tennis games. Eight participants selected Front Row as their number one choice, four participants picked Radio as their top choice, while Television was unanimously ranked as the least preferred option by all participants.



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

Figure 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.

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

7.4 Forced Ranking Results

Figure 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."*

8 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 2.3), 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), sports broadcast accessibility (Section 2.2), and sports video analysis via computer vision (Section 2.3).

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 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 12), and there was no significant difference in feeling of immersion between radio and Front Row (Figure 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 [1]. A direct implication of this is that other racket sports such as badminton, pickleball, table tennis, and squash (collectively shown in Figure 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.

9 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 topspin, 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.

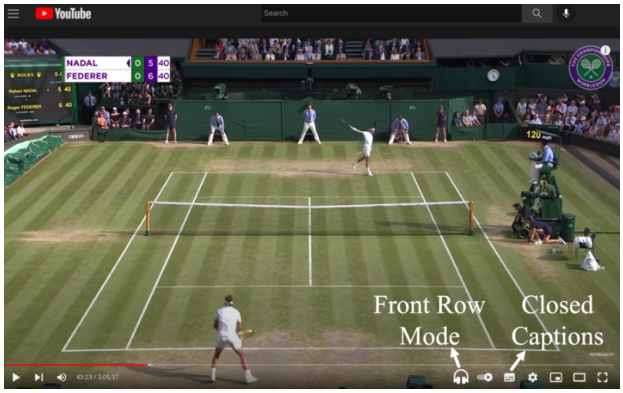


Figure 14: Front Row plug-in for video streaming platforms. Front Row can be integrated with online video streaming platforms, such as YouTube, ESPN+, and Hulu, to make recorded tennis broadcasts accessible to BLV viewers. This could work similarly to how closed captions are implemented on YouTube. Video source: Wimbledon's YouTube channel.

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 [16]. 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 [41] (P4, P5, P8) expressed excitement about sharing their experience of using Front Row with their coaches and friends for this purpose.

10 APPLICATIONS

We now illustrate applications that Front Row could enable in the future. Figure 14 shows a Front Row plug-in for video streaming platforms to make sports videos across the Web accessible. As Figure 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 16 illustrates Front Row's potential in making video games accessible.

11 CONCLUSION

We have presented the Front Row system for automatically generating immersive audio representations of sports from a source 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 equivalently 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.

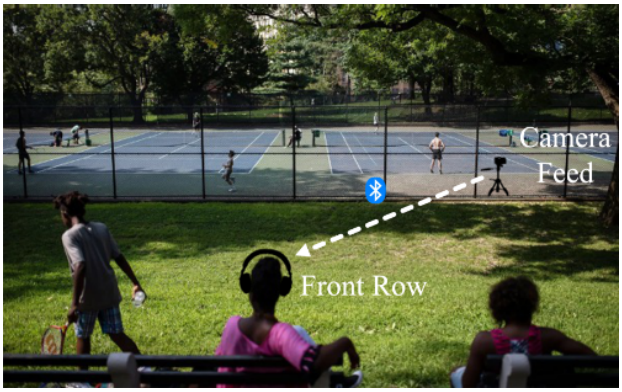


Figure 15: Recreational tennis game at a park. Front Row can make recreational tennis matches, such as matches at high schools, parks, and universities, accessible to BLV viewers. By processing a camera feed captured behind one of the players, Front Row can enable BLV audience members to follow the game in real time. Image source: The New York Times.

Figure 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.

ACKNOWLEDGMENTS

We thank Chloe Ho, Hazel Zhu, Jacqueline Gibson, Jizhong Wang, Siwanta Thapa, and Venkat Ramamoorthi for data annotation, and we thank our study participants for participating in the study. Xin Yi Therese Xu was funded by National Science Foundation Grants 2051053 and 2051060. Conrad Wyrick was supported by the Columbia–Amazon SURE Program.

REFERENCES

[1] Action Audio. 2021. Making Sports Broadcasts Accessible to People Living With Blindness or Low Vision. https://action-audio.com/

[2] American Council of the Blind. 2022. The Audio Description Project. https://adp.acb.org/guidelines.html

[3] Katrin Angerbauer, Nils Rodrigues, Rene Cutura, Seyda Öney, Nelusa Pathmanathan, Cristina Morariu, Daniel Weiskopf, and Michael Sedlmair. 2022. 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*. ACM, New Orleans LA USA, 1–23. https://doi.org/10.1145/3491102.3502133

[4] Saki Asakawa, João Guerreiro, Daisuke Sato, Hironobu Takagi, Dragan Ahmetovic, Desi Gonzalez, Kris M. Kitani, and Chieko Asakawa. 2019. An Independent and Interactive Museum Experience for Blind People. In *Proceedings of the 16th International Web for All Conference*. ACM, San Francisco CA USA, 1–9. https://doi.org/10.1145/3315002.3317557

[5] Saki Asakawa and Amy Hurst. 2021. "What just happened?": Understanding Non-visual Watching Sports Experiences. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Virtual Event USA, 1–3. https://doi.org/10.1145/3441852.3476525

[6] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "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*. ACM, Yokohama Japan, 1–19. https://doi.org/10.1145/3411764.3445498

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Matthew Butler, Leona M Holloway, Samuel Reinders, Cagatay Goncu, and Kim Marriott. 2021. 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 278. https://doi.org/10.1145/3411764.3445207

[9] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Guo Anhong. 2022. OmniScribe: Authoring Immersive Audio Descriptions for 360 ° Videos. (2022), 14.

[10] Zhutian Chen, Shuainan Ye, Xiangtong Chu, Haijun Xia, Hui Zhang, Huamin Qu, and Yingcai Wu. 2022. Augmenting Sports Videos with VisCommentator. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 824–834. https://doi.org/10.1109/TVCG.2021.3114806

[11] Morgan Cottril. February 12, 2020. The Importance of Sports in Culture. https://fghsnews.com/2603/diversity/the-importance-of-sports-in-culture/

[12] N. Dalal and B. Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, San Diego, CA, USA, 886–893. https://doi.org/10.1109/CVPR.2005.177

[13] Expedio Design. 2019. Footbraile. https://www.expediodesign.com/portfolio-footbraille

[14] Olutayo Falase, Alexa F. Siu, and Sean Follmer. 2019. 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*. ACM, Pittsburgh PA USA, 536–538. https://doi.org/10.1145/3308561.3354616

[15] John C. Flanagan. 1954. The critical incident technique. *Psychological Bulletin* 51, 4 (1954), 327–358.

[16] John Garhammer and Harvey Newton. 2013. Applied Video Analysis for Coaches: Weightlifting Examples. *International Journal of Sports Science & Coaching* 8, 3 (Sept. 2013), 581–594. https://doi.org/10.1260/1747-9541.8.3.581 Publisher: SAGE Publications.

[17] Anurag Ghosh and C. V. Jawahar. 2018. SmartTennisTV: Automatic indexing of tennis videos. *arXiv:1801.01430 [cs]* (Jan. 2018). http://arxiv.org/abs/1801.01430 arXiv: 1801.01430.

[18] Anurag Ghosh, Suriya Singh, and C. V. Jawahar. 2017. Towards Structured Analysis of Broadcast Badminton Videos. *arXiv:1712.08714 [cs]* (Dec. 2017). http://arxiv.org/abs/1712.08714 arXiv: 1712.08714.

[19] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3373625.3417027

[20] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376728

[21] Cagatay Goncu and Daniel J. Finnegan. 2021. 'Did You See That!?' Enhancing the Experience of Sports Media Broadcast for Blind People. In *Human-Computer Interaction – INTERACT 2021*. Vol. 12932. Springer International Publishing, Cham, 396–417. https://doi.org/10.1007/978-3-030-85623-6_24

[22] Cagatay Goncu, Anuradha Madugalla, Simone Marinai, and Kim Marriott. 2015. Accessible On-Line Floor Plans. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Florence Italy, 388–398. https://doi.org/10.1145/2736277.2741660

[23] Cagatay Goncu and Kim Marriott. 2011. GraVVITAS: Generic Multi-touch Presentation of Accessible Graphics. In *Human-Computer Interaction – INTERACT 2011*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes,

Philippe Palanque, and Marco Winckler (Eds.). Vol. 6946. Springer Berlin Heidelberg, Berlin, Heidelberg, 30–48. https://doi.org/10.1007/978-3-642-23774-4_5

[24] Leo A. Goodman. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32, 1 (1961), 148–170. https://www.jstor.org/stable/2237615

[25] Alex Graves. 2014. Generating Sequences With Recurrent Neural Networks. http://arxiv.org/abs/1308.0850 arXiv:1308.0850 [cs].

[26] Giles Hamilton-Fletcher, Marianna Obrist, Phil Watten, Michele Mengucci, and Jamie Ward. 2016. "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*. ACM, San Jose California USA, 2162–2174. https://doi.org/10.1145/2858036.2858241

[27] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780. Publisher: MIT press.

[29] Leona M Holloway, Cagatay Goncu, Alon Ilsar, Matthew Butler, and Kim Marriott. 2022. Infosonics: Accessible Infographics for People who are Blind using Sonification and Voice. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–13. https://doi.org/10.1145/3491102.3517465

[30] Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí İk, and Wen-Chih Peng. 2019. 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)*. 1–8. https://doi.org/10.1109/AVSS.2019.8909871 ISSN: 2643-6213.

[31] Kenneth A. Hunt, Terry Bristol, and R. Edward Bashaw. 1999. A conceptual approach to classifying sports fans. *The Journal of Services Marketing* 13, 6 (1999), 439–452. https://doi.org/10.1108/08876049910298720

[32] Hawk-Eye Innovations. 2001. https://www.hawkeyeinnovations.com/

[33] IrisVision. Retrieved July 15, 2022. IrisVision. https://irisvision.com/product/

[34] Hiroo Iwata, Hiroaki Yano, Fumitaka Nakaizumi, and Ryo Kawamura. 2001. Project FEELEX: Adding Haptic Surface to Graphics. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '01*. ACM Press, Not Known, 469–476. https://doi.org/10.1145/383259.383314

[35] Gaurav Jain, Basel Hindi, Connor Courtien, Conrad Wyrick, Xin Yi Therese Xu, Michael C Malcolm, and Brian A. Smith. 2023. Towards Accessible Sports Broadcasts for Blind and Low-Vision Viewers. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–7. https://doi.org/10.1145/3544549.3585610

[36] Grant Jarvie, James Thornton, and Hector Mackie. 2017. *Sport, Culture and Society: An Introduction* (3 ed.). Routledge, Third edition. | Abingdon, Oxon ; New York, NY : Routledge is an imprint of the Taylor & Francis Group, an Informa Business, [2017].

[37] Glenn Jocher et al. April 2021. YOLOv5. https://ultralytics.com/yolov5

[38] Shaun K. Kane, Meredith Ringel Morris, and Jacob O. Wobbrock. 2013. Touchplates: Low-Cost Tactile Overlays for Visually Impaired Touch Screen Users. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Bellevue Washington, 1–8. https://doi.org/10.1145/2513383.2513442

[39] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3501966

[40] Franklin Mingzhe Li, Lotus Zhang, Maryam Bandukda, Abigale Stangl, Kristen Shinohara, Leah Findlater, and Patrick Carrington. 2023. Understanding Visual Arts Experiences of Blind People. https://doi.org/10.1145/3544548.3580941 arXiv:2301.12687 [cs].

[41] Thomas Lin. 2012. Hitting the Court, With an Ear on the Ball. *The New York Times* (June 2012). https://www.nytimes.com/2012/06/05/science/a-game-of-tennis-tests-notions-of-blindness.html

[42] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445233

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 3431–3440. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[44] Kelly Mack, Danielle Bragg, Meredith Ringel Morris, Maarten W. Bos, Isabelle Albi, and Andrés Monroy-Hernández. 2020. Social App Accessibility for Deaf Signers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–31.

[45] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3441852.3471207

[46] Anuradha Madugalla, Kim Marriott, Simone Marinai, Samuele Capobianco, and Cagatay Goncu. 2020. Creating Accessible Online Floor Plans for Visually Impaired Readers. *ACM Transactions on Accessible Computing* 13, 4 (Oct. 2020), 1–37. https://doi.org/10.1145/3410446

[47] J. Matas, C. Galambos, and J. Kittler. 1998. Progressive Probabilistic Hough Transform. In *Procedings of the British Machine Vision Conference 1998*. British Machine Vision Association, Southampton, 26.1–26.10. https://doi.org/10.5244/C.12.26

[48] Tom McEwan and Ben Weerts. 2007. ALT Text and Basic Accessibility. https://doi.org/10.14236/ewic/HCI2007.64

[49] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–11. https://doi.org/10.1145/3173574.3173633

[50] Vishnu Nair, Jay L Karp, Samuel Silverman, Mohar Kalra, Hollis Lehv, Faizan Jamil, and Brian A. Smith. 2021. NavStick: Making Video Games Blind-Accessible via the Ability to Look Around. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 538–551. https://doi.org/10.1145/3472749.3474768

[51] Yuri Nishikawa, Hitoshi Sato, and Jun Ozawa. 2018. Multiple sports player tracking system based on graph optimization using low-cost cameras. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*. 1–4. https://doi.org/10.1109/ICCE.2018.8326126 ISSN: 2158-4001.

[52] NVivo. 1997. NVivo. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[53] Hiroyuki Ohshima, Makoto Kobayashi, and Shigenobu Shimada. 2021. 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*. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451737

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[55] Peter Meijer. Retrieved August 2022. The vOICe. https://www.seeingwithsound.com/

[56] Bridget Pettitt, Katharine Sharpe, and Steven Cooper. 1996. AUDETEL: Enhancing television for visually impaired people. *British Journal of Visual Impairment* 14, 2 (May 1996), 48–52. https://doi.org/10.1177/026461969601400202 Publisher: SAGE Publications Ltd.

[57] Venkatesh Potluri, Tadashi E Grindeland, Jon E. Froehlich, and Jennifer Mankoff. 2021. Examining Visual Semantic Understanding in Blind and Low-Vision Technology Users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445040

[58] Denise Prescher, Jens Bornschein, Wiebke Kohlmann, and Gerhard Weber. 2018. Touching Graphical Applications: Bimanual Tactile Interaction on the HyperBraille Pin-Matrix Display. *Universal Access in the Information Society* 17, 2 (June 2018), 391–409. https://doi.org/10.1007/s10209-017-0538-8

[59] Arthur A Raney and Jennings Bryant. 2006. *Handbook of Sports and Media*. Chapter 19: Why we watch and enjoy mediated sports.

[60] Kyle Rector, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Ringel Morris. 2017. 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* 1, 3 (Sept. 2017), 1–21. https://doi.org/10.1145/3130958

[61] Andreas Reichinger, Stefan Maierhofer, and Werner Purgathofer. 2011. High-Quality Tactile Paintings. *Journal on Computing and Cultural Heritage* 4, 2 (Nov. 2011), 1–13. https://doi.org/10.1145/2037820.2037822

[62] Santander. 2019. Fieeld. https://www.santander.com/en/press-room/press-releases/santander-presents-fieeld-a-deviceenabling-blind-people-to-watch-football-using-their-fingertips

[63] Ather Sharif, Olivia H. Wang, Alida T. Muongchan, Katharina Reinecke, and Jacob O. Wobbrock. 2022. VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517431

[64] Roy Shilkrot, Jochen Huber, Connie Liu, Pattie Maes, and Suranga Chandima Nanayakkara. 2014. FingerReader: a wearable device to support text reading on the go. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 2359–2364. https://doi.org/10.1145/2559206.2581220

[65] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. 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. ACM, Seoul Republic of Korea, 2363–2372. https://doi.org/10.1145/2702123.2702421

[66] Jaeeun Shin, Jundong Cho, and Sangwon Lee. 2020. 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*. ACM, Honolulu HI USA, 1–7. https://doi.org/10.1145/3334480.3383003

[67] Alexa Siu, Gene S-H Kim, Sile O'Modhrain, and Sean Follmer. 2022. Supporting Accessible Data Visualization Through Audio Data Narratives. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517678

[68] Alexa F. Siu, Son Kim, Joshua A. Miele, and Sean Follmer. 2019. 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*. ACM, Pittsburgh PA USA, 342–354. https://doi.org/10.1145/3308561.3353782

[69] Brian A. Smith and Shree K. Nayar. 2018. 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 (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174090

[70] Joel Snyder. 2005. Audio description: The visual made verbal. *International Congress Series* 1282 (Sept. 2005), 935–939. https://doi.org/10.1016/j.ics.2005.05.215

[71] Nancy Staggers and David Kobus. 2000. 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* 7, 2 (2000), 164–176. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC61470/

[72] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404

[73] Lee Stearns, Victor DeSouza, Jessica Yin, Leah Findlater, and Jon E. Froehlich. 2017. 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*. ACM, Baltimore Maryland USA, 361–362. https://doi.org/10.1145/3132525.3134812

[74] Takamasa Tsunoda, Yasuhiro Komori, Masakazu Matsugu, and Tatsuya Harada. 2017. Football Action Recognition Using Hierarchical LSTM. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Honolulu, HI, USA, 155–163. https://doi.org/10.1109/CVPRW.2017.25

[75] Valve Corporation. 2018. Steam Audio. https://valvesoftware.github.io/steam-audio/

[76] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 3866–3874. https://doi.org/10.1109/CVPRW50498.2020.00450

[77] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445347

[78] Yanan Wang, Ruobin Wang, Crescentia Jung, and Yea-Seul Kim. 2022. 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*. ACM, New Orleans LA USA, 1–20. https://doi.org/10.1145/3491102.3517469

[79] World Wide Web Consortium (W3C). 2022. Making Audio and Video Media Accessible. https://www.w3.org/WAI/media/av/

[80] World Wide Web Consortium (W3C). 2022. W3C Image Concepts. https://www.w3.org/WAI/tutorials/images/

[81] Bosun Xie. 2013. *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. J. Ross Publishing. Google-Books-ID: fvDLCgAAQBAJ.

[82] Mingrui Ray Zhang, Mingyuan Zhong, and Jacob O. Wobbrock. 2022. Ga11y: An Automated GIF Annotation System for Visually Impaired Users. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3502092