Low-Cost Vision-Based Multi-Person Foot Tracking for CAVE Systems with Under-Floor Projection

Sebastian Freitag, Sebastian Schmitz, Benjamin Weyers, Torsten W. Kuhlen

Virtual Reality Group
Seffenter Weg 23
52074 Aachen
Tel.: +49 (0)241 / 80 - 24783
E-Mail: freitag@vr.rwth-aachen.de

Abstract: In this work, we present an approach for tracking the feet of multiple users in CAVE-like systems with under-floor projection. It is based on low-cost consumer cameras, does not require users to wear additional equipment, and can be installed without modifying existing components. If the brightness of the floor projection does not contain too much variation, the feet of several people can be successfully and precisely tracked and assigned to individuals. The tracking data can be used to enable or enhance user interfaces like Walking-in-Place or torso-directed steering, provide audio feedback for footsteps, and improve the immersive experience for multiple users.

Keywords: Foot Tracking, CAVE, Cameras, Low-Cost, Under-Floor Projection

1 Introduction

Accurate user tracking is an important prerequisite for many Virtual and Augmented Reality applications. Detecting position and orientation of the user's head is usually the minimum requirement, as it is necessary for a viewer-centered projection. In most cases, this is achieved using marker-based opto-electronic or electromagnetic tracking systems. They also allow accurate tracking of additional features of a user's pose, like her hands and feet, or input devices like a wand or Flystick. However, extending the tracking this way is usually a trade-off decision, as more tracking markers or devices are often intrusive or inconvenient for the user, which can reduce presence and user acceptance.

As a non-intrusive alternative, vision-based tracking systems like the Microsoft Kinect[1] or the Leap Motion[2] can be used. However, these systems have a limited field of view and range, and suffer from occlusion problems due to the single viewpoint, as well as a limited precision. While it is possible to combine several of them to achieve full-body tracking [BKKF13], it is not trivial to do so in CAVE-like virtual environments without blocking any of the screens.

[1] www.xbox.com/kinect
[2] www.leapmotion.com

However, tracking only certain features instead of the complete body pose, such as feet, can lead to precise results with limited hardware expenditure, and still provides additional information. We present an approach for multi-user foot tracking in CAVE-like virtual environments with under-floor projection, which builds upon the fact that foot shadows can usually be seen from below the floor projection plane.

Systems based on back-projected floor planes that are able to accurately track feet have already been presented in the past (e.g., [AKM+10, BHH+13]). These are typically based on FTIR [Han05], which allows accurate and highly resolved recognition over the whole area, making tracking of objects and users in different poses possible as well [BHH+13, SFK+15]. However, this technique requires infrared light sources shining into the floor plane from the side, which usually constitutes a non-trivial and possibly costly addition to systems not already equipped with these. In contrast, our approach can be added to existing CAVEs with under-floor projection easily, relying only on low-cost consumer hardware.

Feet can also be tracked using markers or sensors (e.g., [FWW08, WWBJ10]). However, these systems require users to wear additional hardware devices and possibly also need a user calibration, both of which is not necessary in our approach. In CAVE systems where users have to wear protective slippers, these can be equipped with markers for the already existing tracking system. However, CAVE tracking systems are often optimized for the typical interaction regions between waist and head height, providing low tracking quality close to the floor. Additionally, for opto-electronic tracking with the tracking cameras usually mounted above, occlusion through other body parts is an issue.

Although tracking a user's foot shadows in under-floor projection systems visually has been previously explored [ZMB11], only a best-case scenario with a single user, a single camera and a nearly black floor projection without artifacts was considered.

In addition to enabling user interfaces like Walking-In-Place [ZMB11] and foot gestures (e.g., [JFKZ01]), the results can be used to provide audio feedback for footsteps [MCV+13]. Furthermore, as the user's body posture can be estimated from her feet positions, interfaces like torso-directed steering [BKH98] can be implemented without additional tracking targets. Moreover, the position information can be used to correct conflicting depth cues in a multi-user setting (e.g., by removing or cutting out objects) and adapt the content for secondary users, e.g., by showing additional content, or moving annotations. Moreover, they provide more cues about the pose of all head-tracked users—e.g., if they are leaning over or sideways—which can be used for more precise interfaces and simulations.

The paper is structured as follows. Section 2 introduces the proposed method, including the main tracking procedure and an alternative approach for unfavorable lighting situations, and discusses limitations of both techniques. Section 3 presents a preliminary study of the efficiency and precision of our implementation, along with a qualitative evaluation of the accuracy. Finally, section 4 concludes the paper and provides an outlook on future work.


Figure 1: **Left:** Five users in a virtual scene, seen from below the under-floor projection system. The shadows of all feet are visible. **Right:** Photograph of feet in favorable conditions on the projection of a textured carpet. The white artifacts are specular reflections from the projectors.

2 Method

Our approach is based on the observation that in under-floor projection systems, the users' feet can be seen from below as shadows on the projection screen (cf. Figure 1). We detect these shadows from images captured by consumer cameras, and then determine foot positions and orientations. In the following, we allow for users wearing protective slippers on their feet, which is often the case in CAVE environments. These are usually symmetrical regarding their front and heel side, which, in contrast to actual feet or shoe shadows that have a more pronounced shape, slightly complicates the orientation estimation, as front and back cannot trivially be differentiated (cf. Figure 1).

2.1 Setup

In our setup (illustrated in Figure 2), we used two consumer webcams (Logitech HD Pro Webcam C920) connected to a regular Windows PC via USB. The reasons for using two cameras are that in our case, the field of view of a single webcam is insufficient to cover all of the projection screen, and a beam in the center of the floor occludes part of the screen from any point of view. However, the tracking quality can be further improved with more cameras to increase stability and cope with artifacts such as specular reflections from the projectors (cf. Figure 1, right and Figure 4, left).

Our prototype tracking application is implemented in C++, using image processing algorithms from the open source computer vision library OpenCV[3].

[3] www.opencv.org

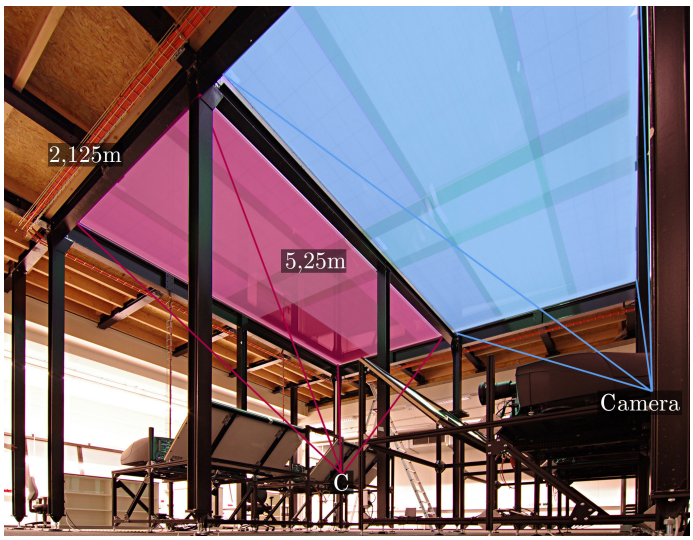


Figure 2: Images were obtained by mounting two Logitech HD Pro C920 consumer webcams (≈65 €) below our CAVE. A specialty of this system is that the floor is partitioned by a sustaining beam. To avoid occlusion, this makes the use of at least two cameras necessary which cannot be placed below the center of the floor.

2.2 Main Procedure

First, images are taken from all cameras at the same time. Then, the following processing steps are applied:

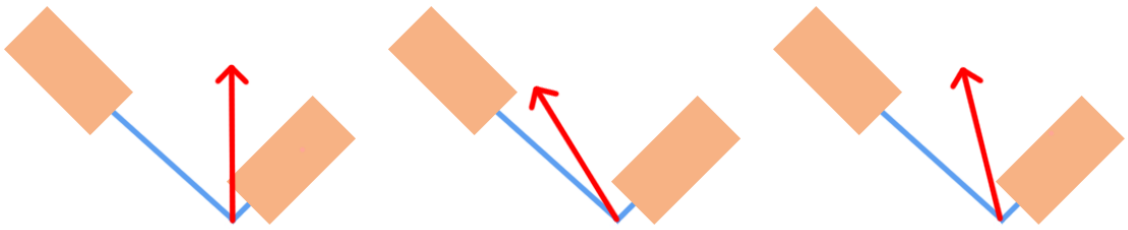
1. **Image transformation**: A perspective transformation is applied to the images, and they are stitched and cropped to produce a single image providing an orthogonal view of the projection screen. The perspective transformation has to be defined once for a camera setup, which is done by a user clicking on the corners of the projection screen in the camera images using a simple calibration tool included in our tracking application.
2. **Smoothing**: The image is converted to grayscale (feet shadows mainly differ from the rest of the image in brightness, but not color) and smoothed with a Gaussian filter. The smoothing reduces noise and artifacts, but empirically does not change the (usually fuzzy) foot shadows much.
3. **Binarization**: The image is binarized using a variant of Otsu's method [Ots75] that accounts for the fact that foot shadows only cover a very small portion of the total image and the projected image usually contains a wide range of brightness values.
4. **Enhancement**: The binarized image is processed with morphological operations (opening, closing). This removes small-scale clutter and closes shadows that are partially split by brighter stripes or reflections (cf. Figure 4, left). Furthermore, two foot shadows that are connected by only a small juncture are separated.
5. **Edge detection**: A Canny edge detection [Can86] is performed to find closed boundaries. All boundaries that represent dark shapes on a lighter background are stored as possible candidates for foot shadows. Note that it is equivalent to detect connected

components and continue the analysis on the pixel image, although the candidate description would not be as compact at this point.

6. **Rectangle representation**: For every candidate, the minimum oriented 2D bounding box is computed. This provides an estimate for possible feet lengths, widths, and rotations.
7. **Feet detection**: All boxes corresponding in length and width to an empirically determined range of 60–180% of the expected foot size are selected as feet. We observed that this range excludes most unintended clutter and avoids the detection of two feet standing very close together as one foot, but still allows for some variation in the shadow caused by users' legs, or feet that do not completely touch the floor (e.g., in motion). Note that if equally sized protective slippers are used, the range can be restricted more than when foot shadows have different sizes.

   As midpoint of each foot, the center of mass of its contour is used. The rotation is extracted from the oriented bounding box, where the two shorter sides are interpreted as front and back. However, at this point, it is not decided which of these is which.
8. **Feet assignment**: Two feet each are assigned to a person. For this, the feet's last assignment (if available) as well as expected step lengths [Kuo01] and distances between feet of the same person are considered, also allowing feet to be lifted for short periods of time without changing the association to a person. The assignment can be formulated as a linear optimization problem minimizing distances between feet of the same person, although we found that first assigning unambiguous feet (that only have one possible partner) and then greedily selecting for distance, works well in all realistic situations. Note that to ensure a correct assignment of feet to persons, all users should enter the system with both feet. However, this step allows for the tracking of single feet to be lost for a while to account for users lifting their feet or entering the system one foot at a time. When tracking of a foot is lost (e.g., because it was lifted), already assigned feet are only reassigned after five seconds. Assignments of new feet are only made if the number of unassigned feet is even.
9. **Orientation**: Up to this point, it is not clear which of the shorter sides of each foot rectangle is the front-facing one, and if a foot is a left or right one. However, most people tend to stand in a V-shaped manner most of the time [MM97] with an average opening angle of 14°, the open end of the V being the front side. From this, the full orientation can be deduced and is retained as long as the observation does not continually indicate otherwise.
10. **Person posture estimation**: The position of a user's feet provides additional information about her body posture. Usually, her torso points in approximately the same direction as her feet, which can be used in algorithms relying on the orientation of the upper body (e.g., torso-directed steering [BKH98]) without having to equip the user with additional tracking markers.

    To facilitate the use of these methods, we compute the person's position (an approxi-



(a) (b) (c)

Figure 3: Possible approaches to determine a person's overall body orientation based on foot rotations. The blue lines visualize the foot directions, meeting in the *intersection point*. (a) Average of foot angles. (b) Line between *intersection point* and feet midpoint. (c) Combined foot angles, weighted by distance between foot and *intersection point*.

mation of her 2D center of mass) as the midpoint between the centers of mass of both feet. Furthermore, the person's orientation can be determined from her feet orientations. We developed three different approaches for this (cf. Figure 3), of which the best-fitting one has to be determined in a user study in future work.

The results are foot positions, orientations and assignments to individual persons, as well as the persons' position and rotation information for arbitrarily many users. Furthermore, users can also enter and leave the system at any time, as no calibration or initialization is necessary. These data are sent to the running application over a network interface.

2.3 Alternative Approach based on Foreground-Background Separation

The process as described above works best for floor projections with (more or less) uniform brightness. However, when there are large differences in brightness (e.g., a checkerboard pattern), foot shadows on the bright areas are actually much brighter than the dark areas without shadows—at least in our system—and less contrasted than elsewhere due to the strong light from below. Thus, in these situations, feet are only correctly recognized on the dark areas. Furthermore, we observed some color-texture combinations where foot shadows almost completely disappeared (cf. Figure 4, right).

However, we observed that even when foot shadows disappeared almost completely, they were still visible in motion. Therefore, we integrated a second approach, based on Zivkovic' foreground-background separation algorithm [ZvdH06]. The main idea behind this approach is to train a model of the "background" (the projection without the foot shadows), and extract the "foreground" (the foot shadows) from the difference of the captured image and the background model.

In order to obtain a model of the background, there are different possible methods. As the image that is projected on the floor is completely generated by the running application anyway, it is possible to directly obtain the original background image. One way to do

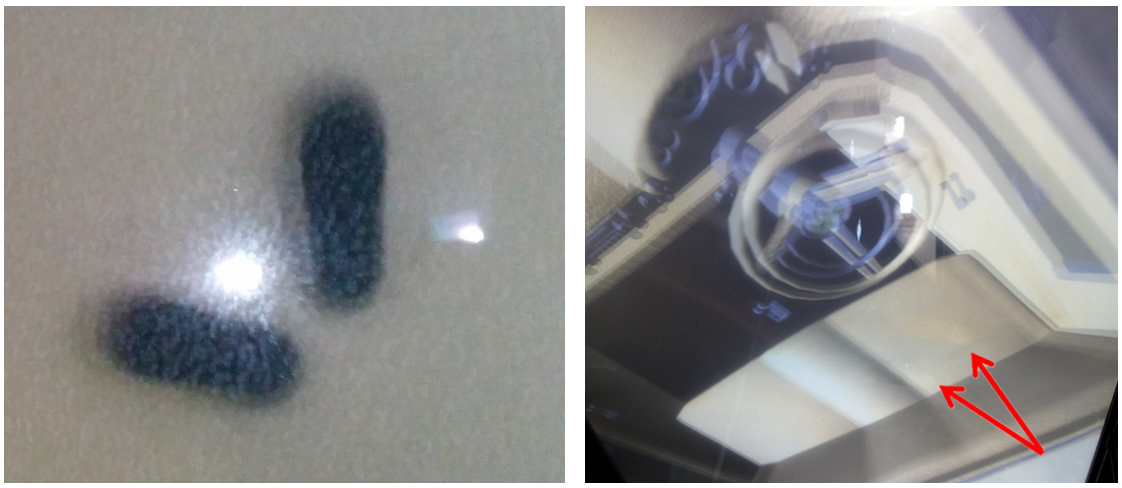


Figure 4: **Left:** Specular reflections from the projectors can obfuscate the foot shapes when the feet are positioned above. **Right:** For some projections—here, the user is standing inside a car—the feet (indicated by red arrows) are nearly impossible to find in a static image.

this would be to have the corresponding graphics nodes send the image to the analysis PC via a network connection. However, this would not only cause massive network load, but also additional overhead on the corresponding graphics node(s). Another way would be to generate the floor projection image directly on the analysis PC. However, this would require a synchronization of the analysis PC with the rest of the visualization cluster that drives the CAVE application, and therefore an integration with the CAVE system. In addition to requiring an analysis PC that is at least as powerful as the CAVE's graphics nodes, it would reduce the flexibility and loose coupling of the approach.

Therefore, we opted for creating and updating the background model directly from the last camera frames. With a largely static floor projection, this allows to capture feet in movement even when it is very difficult to detect them in a single image. However, as users usually only move one foot at a time, the stationary foot migrates into the background model after a short time. Therefore, we only estimate the user's body position from the position of the foot, averaging over several frames to account for the feet of the user moving in alternation.

To ensure a good tracking quality, the main procedure (section 2.2) is always used by default. If it fails for several frames in sequence, the analysis switches to the foreground-background separation method, but still continuously executes the main procedure to switch back when its results are good enough again.

2.4 Limitations

As already mentioned above, a limitation of the main procedure (section 2.2) are shadows that are not detected if there are large contrast differences in the projected floor image. In many of these cases, however, the shadows could still be detected by their local gradient.

Therefore, a detection using histograms of oriented gradients (HOGs) [DT05] could be used to allow detection in these cases. However, in some cases (cf. Figure 4, right), the approach would still fail, as the foot shadows are just too similar to the rest of the projection. Lastly, an obvious example where the proposed approach must fail in principle is if the floor projection itself contains shapes that look like shadows of feet.

The foreground-background separation approach (section 2.3) only works correctly if the floor projection remains largely static, as all movement is interpreted as possible foot candidates. Although local movement is filtered out as it is usually short-lived and not foot-shaped, when the ground projection changes continually as a whole (e.g., when the user travels virtually), the method fails and tracking is lost (at least if it is not possible to switch back to the main method).

Finally, both approaches require some kind of foot shadow to be visible from below, which requires light from above. However, this is usually provided by the projections on the CAVE walls, as the shadows are discernible even in low-light situations.

Note that these limitations might be avoided when infrared light sources and cameras are available (see section 4).

3 Evaluation

We conducted a preliminary evaluation to assess the efficiency and precision of our implementation. For the processing, we used a low-cost PC with an Intel Core 2 Duo E8400 processor and 2GB RAM running Windows 7. Each method was tested using a 360p (640×360 pixels) and a 720p (1280×720 pixels) image for each of the two cameras, together filming the 5.25m×5.25m CAVE floor (cf. Figure 2). The scenario was a favorable scene featuring a textured, but relatively dark office carpet as ground projection (cf. Figure 1, right).

3.1 Efficiency

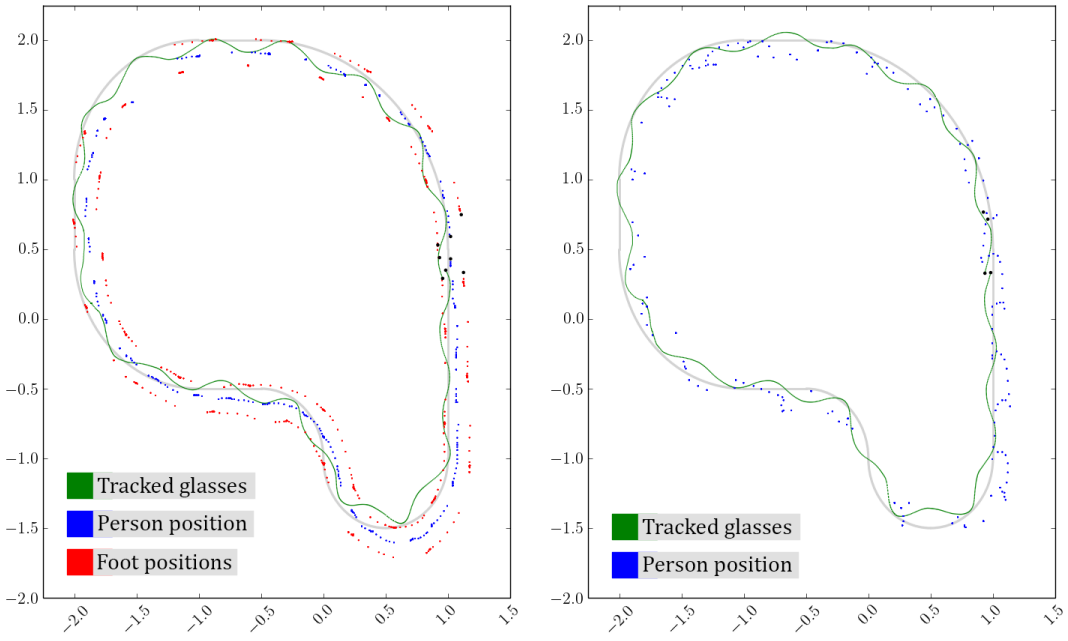
The main procedure (section 2.2) and the foreground-background separation (section 2.3) processed the following number of images per second (averaged over 300 seconds of analysis):

	2×360p	2×720p
main procedure	31.8	9.1
foreground-background separation	16.2	3.7

The performance apparently scales roughly linearly with the number of pixels. It should be noted that with a current processor, it can easily be boosted several times.

3.2 Precision

Four pairs of protective slippers were placed at different positions within the CAVE. For the main procedure, the following percentage of tracking data lay within the given distance from the mean:



(a) Main procedure (b) Foreground-background separation

Figure 5: Qualitative evaluation of foot tracking with participants following a curved route on the ground (indicated by the gray line). The units are meters, the CAVE's walls are at ±2.625m. The first points of the round are displayed as black dots.

	Position		Orientation	
	2×360p	2×720p	2×360p	2×720p
90%	1.7mm	1.5mm	±2.7°	±2.3°
95%	2.1mm	1.9mm	±3.5°	±2.9°
99%	3.2mm	2.7mm	±5.2°	±4.7°

The results show that even the relatively low 360p resolution ($\approx 600 \times 600$ pixels for 5.25m×5.25m floor area) has only subpixel positional jitter. As the scene was static, the deviations are expected to originate from image noise in the low-light environment and the image transformation.

3.3 Qualitative Evaluation of Trajectory Data

As it is difficult to reliably obtain data of the true foot positions to measure the accuracy of the approach quantitatively, we conducted an experiment to qualitatively evaluate the tracking results. In the experiment, participants had to follow a curved line on the ground for several rounds. The tracking data from both the main procedure and the foreground-background separation for a representative participant of the third round (to exclude noise from the participant getting used to the environment) are visualized in Figure 5.

The foot positions tracked by the main procedure (Figure 5a, red) correspond to the expectation of a person walking on the indicated route, with both feet in about the same

Figure 6: **Left:** Color photograph of the CAVE floor showing a problematic scene. **Right:** Enhanced infrared capture of the same scene, taken with a Microsoft Kinect. Even though the infrared image is noisy and low resolution, both feet can be distinguished much better than in the color image.

distance from each other most of the time. Both feet, as well as the user's body position, are tracked within short intervals. The regular gaps in the data—where a foot could not be tracked anymore—are caused by users lifting their feet. This is especially noticeable in the straighter sections, where participants walked faster, lifting their feet more. Note that the person's position (blue) is not disturbed by the missing foot information, as it is only updated when both feet are tracked.

For the main procedure, there is a gap in the foot tracking data (Figure 5a, upper left corner), which corresponds to a strong specular reflection from a floor projector at that position, making the foot shadow smaller and harder to recognize. This could be fixed with additional cameras—using four cameras instead of two allows to filter out specular highlights, as each position is observed by at least two cameras. The gap is not visible in the data obtained by the foreground-background separation method (Figure 5b), as parts of the feet in motion can still be distinguished, even when they are nearly directly above the specular highlight. However, the person position information obtained using this method is not as precise as with the main method, as they are only based on the smoothed position of one moving foot at a time.

Furthermore, the data reveal a systematic shift of all tracked foot position in the +x direction, which is probably due to an imperfect calibration of the perspective transformation (section 2.2, step 1).

4 Conclusion

In this work, we presented our approach for low-cost multi-person foot tracking for CAVE-like systems with under-floor projections. In stable conditions with roughly uniform lighting, foot detection and assignment results are precise and efficient; for some scenarios with heterogeneous lighting, at least moving feet are found.

In future work, the important special case of non-uniform projection brightness will be approached. The foot detection in this case could be improved by incorporating gradient-based methods (e.g., [DT05]), instead of only relying on brightness information.

Furthermore, a promising alternative would be the combination with an image from an infrared camera, which, in contrast to visible light images, would not be affected by the projector image. Many CAVE systems use infrared-based optoelectronic tracking, which provides an already existing source of infrared light from above the CAVE floor. Even though the infrared light from most tracking cameras is very weak when seen through a projection screen, it may be sufficient with appropriate post-processing. As a low-cost approach to this problem, consumer cameras like the one built into the Microsoft Kinect can be used (see Figure 6 for an image taken with a Kinect camera, strongly increased in brightness).

We also plan to study how accurate the body position and orientation can be estimated from tracking the feet, and how novel user interfaces can benefit from the additional information. Furthermore, we want to investigate how a multi-person immersive experience can be enhanced when all users are tracked without requiring special equipment.

References

[AKM+10] T. Augsten, K. Kaefer, R. Meusel, C. Fetzer, D. Kanitz, T. Stoff, T. Becker, C. Holz, and P. Baudisch. Multitoe: High-Precision Interaction with Back-Projected Floors based on High-Resolution Multi-Touch Input. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 209–218, 2010.

[BHH+13] A. Bränzel, C. Holz, D. Hoffmann, D. Schmidt, M. Knaust, P. Lühne, R. Meusel, S. Richter, and P. Baudisch. GravitySpace: Tracking Users and their Poses in a Smart Room using a Pressure-Sensing Floor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 725–734, 2013.

[BKH98] Doug A. Bowman, David Koller, and Larry F. Hodges. A Methodology for the Evaluation of Travel Techniques for Immersive Virtual Environments. *Virtual Reality*, 3(2):120–131, 1998.

[BKKF13] Stephan Beck, Andre Kunert, Alexander Kulik, and Bernd Froehlich. Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013.

[Can86] John Canny. A Computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[DT05] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[FWW08] Jeff Feasel, Mary C. Whitton, and Jeremy D. Wendt. LLCM-WIP: Low-Latency, Continuous-Motion Walking-in-Place. In *IEEE Symposium on 3D User Interfaces*, pages 97–104, 2008.

[Han05] Jefferson Y Han. Low-Cost Multi-Touch Sensing through Frustrated Total Internal Reflection. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 115–118. ACM, 2005.

[JFKZ01] Joseph J. LaViola Jr., Daniel Acevedo Feliz, Daniel F. Keefe, and Robert C. Zeleznik. Hands-Free Multi-Scale Navigation in Virtual Environments. In *Proc. of the 2001 Symp. on Interactive 3D Graphics, SI3D*, pages 9–15, 2001.

[Kuo01] Arthur D. Kuo. A Simple Model of Bipedal Walking Predicts the Preferred Speed–Step Length Relationship. *Journal of Biomechanical Engineering*, 123(3):264–269, 2001.

[MCV+13] Maud Marchal, Gabriel Cirio, Yon Visell, Federico Fontana, Stefania Serafin, Jeremy Cooperstock, and Anatole Lécuyer. Multimodal Rendering of Walking over Virtual Grounds. In *Human Walking in Virtual Environments*, pages 263–295. Springer, 2013.

[MM97] W.E. McIlroy and B.E. Maki. Preferred Placement of the Feet during Quiet Stance: Development of a Standardized Foot Placement for Balance Testing. *Clinical Biomechanics*, 12(1):66–70, 1997.

[Ots75] Nobuyuki Otsu. A Threshold Selection Method from Gray-level Histograms. *Automatica*, 11(285-296):23–27, 1975.

[SFK+15] D. Schmidt, J. Frohnhofen, S. Knebel, F. Meinel, M. Perchyk, J. Risch, J. Striebel, J. Wachtel, and P. Baudisch. Ergonomic Interaction for Touch Floors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3879–3888, 2015.

[WWBJ10] Jeremy D. Wendt, Mary C. Whitton, and Frederick P. Brooks Jr. GUD WIP: Gait-Understanding-Driven Walking-In-Place. In *IEEE Virtual Reality Conference*, pages 51–58, 2010.

[ZMB11] David J. Zielinski, Ryan P. McMahan, and Rachael B Brady. Shadow Walking: An Unencumbered Locomotion Technique for Systems with Under-Floor Projection. In *IEEE Virtual Reality Conference*, pages 167–170, 2011.

[ZvdH06] Zoran Zivkovic and Ferdinand van der Heijden. Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction. *Pattern recognition letters*, 27(7):773–780, 2006.