Measurement of Feet-Ground Interaction for Real-Time Person Localization, Step Characterization, and Gaming

Chris Varekamp
chris.varekamp@philips.com

Patrick Vandewalle
patrick.vandewalle@philips.com

Philips Research, High Tech Campus 36
5656AE Eindhoven, The Netherlands

ABSTRACT

We introduce a setup for detailed and unobtrusive feet mapping in a room. Two cameras arranged in a stereo pair detect when an object touches ground by analyzing the occlusions on a fluorescent tape that is attached to the baseboard of a room. The disparity between both cameras allows localization of a person's feet and the calculation of step size and walking speed. People are separated from furniture such as chairs and tables by studying the occlusion duration. We present and discuss data-association and filtering algorithms and the algorithms that are needed for presence detection, step characterization and gaming.

ACM Classification Keywords

H5.1.b Artificial, augmented, and virtual realities; H.5.m. Information interfaces and representation: Miscellaneous, I.2.10.a 3D/stereo scene analysis

INTRODUCTION

Our aim is to create a low-cost measurement setup for long-term installation in homes such that walking behavior can be studied on natural home surfaces or gaming can be done in-home. We introduce a measurement setup that allows unobtrusive measurement for this purpose. The setup consists of two cameras (stereo pair) on one side and an elongated fluorescent tape on the other side of a room. Both the cameras and the tape are mounted near the floor. The tape height is only 2 cm to be minimally disturbing someone that is present in the room. The measurement geometry is shown in Figure 1. The setup does not require on-body sensors or markers. We calculate and visualize in real-time the 2D position of feet and other objects that touch the floor. When observed over time, this measurement provides information on the walking behavior of people. Our measurement setup has potential applications ranging from presence detection to in-home gaming. Most existing systems that measure feet-ground interaction are intended for professional use, such as systems that measure gait. We are interested in in-home activity monitoring using for instance step size as an indicator and not for clinical research. The gaming industry is working on unobtrusive solutions for gaming. Notably, Microsoft plans to use a depth sensor such that a game controller is no longer required [1].

This allows for the use of gestures, and also for kicking a ball as we will also demonstrate in this paper. Currently we cannot compare the accuracy of calculated feet positions with those calculated by the Xbox sensor since that sensor is not on the market at the time of writing. Video based human gait extraction typically exploits strong prior shape information. The authors in [2] train a 12 parameter model and fit that to extracted silhouettes. They use strong prior information in an articulated 2D model to compensate for noisy silhouettes (e.g. from background subtraction outdoor). However, they consider only walkers moving perpendicular to the camera. For our soccer game (see further) the player can also face the camera. The use of cameras in combination with a background that has

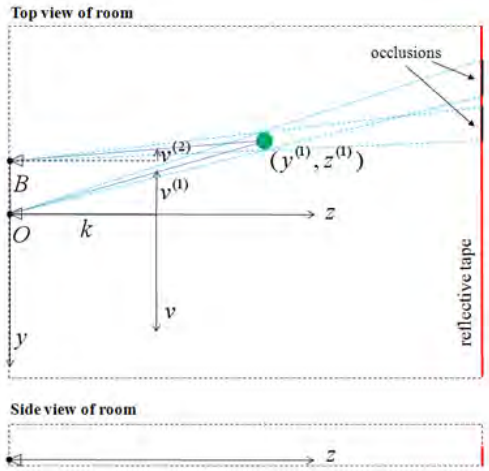

Figure 1. Top view of a room. Two cameras are mounted a distance *B* apart with the origin and camera 1. Image coordinate *v* is assumed to lie parallel to world coordinate *y*. World coordinate *z* points towards the tape. An object that touches ground occludes the reflective tape. The object position $(y^{(1)}, z^{(1)})$ on the ground is calculated from the shift $(v^{(1)} - v^{(2)})$ in the image *v*-coordinate of the occluded area between the two camera images.

specified optical properties is known from the movie industry (blue screen or green screen) and from the monitoring of inventory [3]. We propose a new background based technique for a room that is minimally intrusive and can be used to calculate the 2D position of people and characteristics such as step size and walking speed. In the remainder of this paper we discuss the measurement setup, treat the required signal processing and develop specific application algorithms.

MATERIALS AND METHODS

Measurement Setup

In our experiments, we use two QuickCam Fusion webcams (Logitech). The cameras capture images at a resolution of 640×480 pixels at 30 frames per second. The cameras are placed 15 cm apart and mounted 1.5 cm above the ground. At the opposite side of the room we attach a red fluorescent tape (HPX). The tape is over 3 m long and has a height of only 2 cm. We implement algorithms in C building on data structures, video input/output and plotting functions from the open source computer vision library *OpenCV* originally developed at Intel and now hosted by Willow Garage [4].

Tape Detection

The fluorescent tape serves as a known background against which objects such as human feet are observed. We construct a vector of tape pixels for each camera using a fixed threshold that differs for each color channel (rgb). We use r = {119…255} and g,b = {0…99} as selection criterion for 'tape pixels'. Since the tape is red, the red channel is required to take large values whereas the green and blue channels are required to take small values. In general, the selection of these values depends both on the camera properties, illumination, and on the specific tape that is used. Since the tape touches the ground surface, and the ground surface is flat, the tape pixels lay on a straight line in each camera image. We fit this line to the tape pixels in each image. During this fitting step we clear the carpet of objects. Currently, carpet roughness, which could influence our measurement, is not accounted for.

Object Segmentation

For each tape pixel we detect occluded line segments using again knowledge of the tape color. The start and stop coordinates of each line segment are stored. Figure 2 shows two camera images with the detected line segments that in this case correspond to a foot that occludes the fluorescent tape.

Disparity Estimation

In order to produce a map of a person's feet on the ground we need the disparity of each object between the two cameras. Here we account for the possibility of a false detection due to image noise. If each image contains the same number of objects we associate objects in the two cameras using their order of occurrence along the tape. For each object we calculate the average pixel v-coordinate in both cameras (see the geometry in Figure 1). We then

Figure 2. Left and right stereo images of a a walking person with one foot touching ground. The corresponding occlusion segments of the red fluorescent tape are shown in green below each images of the stereo pair.

calculate disparity $v^{(2)} - v^{(1)}$ of the corresponding points between the two cameras of the stereo pair.

Calculation of Object Position and Size

The observed disparity allows the calculation of depth z. The camera model geometry in Figure 1 is a simplified version of the model given by Forsyth and Ponce [5, p.29-31]. Using this simplified model, the z-coordinate follows as:

$$z^{(1)} = \frac{Bk}{v^{(1)} - v^{(2)}}, \qquad (1)$$

where B is the distance between the cameras [m] and k is a constant that depends on the camera focal length and pixel size. The superscript $^{(1)}$ in the above notation stresses that the origin of our coordinate system lies at the lens centre point of the first camera. When z is known, the y-coordinate follows as:

$$y^{(1)} = \frac{z^{(1)} v^{(1)}}{k}. \qquad (2)$$

We can now create a 2D map showing a top view of the room with all objects that touch the ground. Such a map is shown in Figure 3. The sizes of the circles can differ since one foot can be fully on the ground, while the other foot is in the process of being lifted. We estimate object size from the end points v_{start} and v_{stop} of a line segment:

$$\Delta^{(1)} = \frac{z^{(1)}}{k} \left| v_{\text{start}}^{(1)} - v_{\text{stop}}^{(1)} \right|. \qquad (3)$$

Calibration

Our aim is to be able to map feet positions in a room's coordinate system. We need to determine the baseline B and the camera parameter k. Although B can be known when the cameras are mounted in fixed positions, we consider it unknown since the two cameras may need to be installed

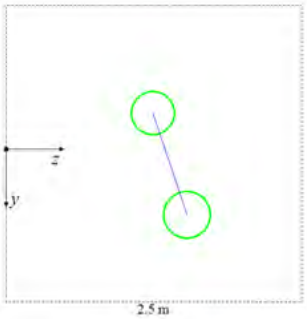

Figure 3. Top view of the room with the position of two feet on the ground. During step analysis, the two feet are recognized as one single step and connected by a line.

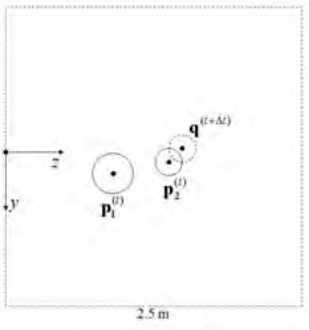

Figure 4. Illustration of the over-time data association of detected positions for the purpose of temporal analysis and filtering.

separately in the baseboard of a room, in which case calibration follows later.

To estimate B and k, black calibration cylinders are placed at known coordinates (y,z). Given these known coordinates, we first solve for parameter k by minimizing the sum of squared errors in the y-coordinate. Using equation (2) the least-squares estimate for k is:

$$\hat{k} = \frac{\sum_{i=1}^{N} z_i^{(1)} v_i^{(1)}}{\sum_{i=1}^{N} y_i^{(1)}}. \qquad (4)$$

Given that k is now known, the baseline B can be calculated after inserting the estimated value for k in equation (1). The least squares estimate for B then follows as:

$$\hat{B} = \frac{\sum_{i=1}^{N} z_i^{(1)}}{\hat{k} \sum_{i=1}^{N} \left(\frac{1}{v_i^{(1)} - v_i^{(2)}} \right)}. \qquad (5)$$

Note that the above least-squares method is sub-optimal since it does not minimize the Euclidian distance between the true and estimated (y,z)–coordinates. However, the above approach that treats each coordinate separately has the advantage of providing a closed-form solution.

Over-Time Association and Filtering

We link feet measurements over time to remove noise. This temporal linking is also necessary to determine when a foot is placed on the ground and when it is lifted. Data association is explained in Figure 4. Let $\mathbf{p}$ denote the 2D location of an object. Now assume that two objects are present at time t and that one new measurement $\mathbf{q}$ is detected at time $t+\Delta t$. The new measurement $\mathbf{q}$ lies close to $\mathbf{p}_2$ and far from $\mathbf{p}_1$. We use the new measurement to update and object's location if it lies less than 10 cm from that object. Thus $\mathbf{q}_3$ is used to update $\mathbf{p}_2$ but not to update $\mathbf{p}_1$ (see Figure 4). After this data-association step, $\mathbf{p}_2$ is updated using a low-pass filter. In this case the estimate for $\mathbf{p}_2$ is updated according to:

$$\mathbf{p}_2^{(t+\Delta t)} = \mathbf{p}_2^{(t)} \alpha + \mathbf{q}^{t+\Delta t} (1-\alpha), \qquad (6)$$

where α=0.1 controls the amount of filtering. Note that position $\mathbf{p}_1$ stops to exist since no new measurement has been found in its vicinity. The same filtering operation is also applied to update the size estimate. Currently we do not explicitly treat the feet of multiple people with the consequence that errors occur when multiple persons come close.

Step Detection

During a typical step, first the heel touches ground, then the full foot, and finally when the foot is lifted, the foot leaves the ground via the toes. These events correspond to different spatial positions on the ground. We use a *lead time* of 0.2 s (6 frames on average) to allow the position filter to do its work before we plot the foot position. We have found that this lead time is necessary to avoid introducing multiple positions during the step that do not correspond to a foot touching ground. The occurrence of these non-touching positions could be due to a small clearance distance [6]. The straight line that we fit to the 2 cm high tape will probably lie approximately 1 cm above the ground surface. A smaller clearance is common as shown in [6, p.193, fig.3]. To recognize a step we introduce a minimum step size of 0.3 m. If a detected location has a duration that exceeds the lead time and lies further than 0.3 m from the previous foot location then we recognize it as the next foot of a step.

APPLICATION ALGORITHMS

In this section we develop various application algorithms to show how the measurements must be interpreted for practical use. All application algorithms run in real-time on a conventional laptop.

Person Localization

Presence detection is relevant for home and building

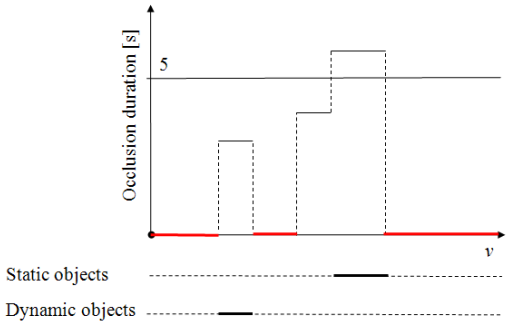

Figure 5. Temporal occlusion analysis as a basis for the separation of dynamic and static objects. One dynamic object is ignored because it has one end-point for which the position cannot be determined.

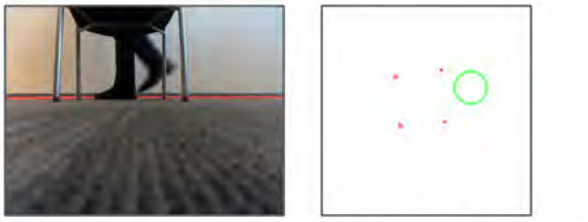

Figure 6. Positioning a person's feet in a room that is cluttered with furniture. Due to occlusion a foot position may be missed. However, if detected, the positions are not influenced by the presence of static objects.

automation such as lighting control. We wish to position people in a room that is possibly cluttered by tables and/or chairs. With the processing discussed so far, this can be problematic. A person's foot placed behind a chair can result in a single silhouette instead of two separate ones and the estimated position will lie somewhere between the foot and the chair. To solve this problem we store the duration in seconds that a 'tape pixel' has already been occluded. We then produce two object segmentations: one containing occlusions with duration longer than 5 seconds and another that consists of occlusions with duration of less than 5 seconds (Figure 5). Each segmentation is input to position/size estimation and data-association/filtering. Figure 6 shows an example classification where a person passes behind a chair.

Soccer

We implemented a real-time soccer game. To realize a virtual ball-kick, the velocity vector of each foot is calculated. To increase accuracy we exploit the typical smooth leg motion and filter the velocity vector temporally:

$$\mathbf{v}^{(t+\Delta t)} = \mathbf{v}^{(t)}\beta + (\mathbf{p}^{(t+\Delta t)} - \mathbf{p}^{(t)})(1-\beta)\,. \qquad (7)$$

where β=0.5 weighs the contribution from the previous foot motion vector $v^{(t+\Delta t)}$. With this filtering operation we achieve a realistic kick. If one of the two feet hits the virtual ball, we transfer the foot-velocity vector to the ball and let it move according to the motion vector at the time of contact (Figure 7).

Tap Dance

We pre-record sounds that resemble the effect of a *heel* hitting a wooden floor and a different sound for a *toe* hitting a wooden floor. The sounds play in real-time in response to the feet actions. To determine whether heal or toe hit the ground, we compare the *z*-coordinate of each foot when it hits the ground with the mean *z*-coordinate of both feet that we track over time. If the tap dancer faces roughly in the negative *z*-direction, the classification works satisfactorily.

Figure 7. Kicking a virtual soccer ball (green circle) with one of two feet (blue circle) during a simple soccer play.

CONCLUSIONS AND FUTURE WORK

We have introduced a low-cost setup for measuring ground-feet dynamics. The technique can determine presence, foot location, step size, walk direction and speed and can be meaningful for in-home gaming.

The setup was tested using an image resolution of 640×480. The image resolution could be further increased to achieve higher accuracy of feet maps. The proposed measurement system has a number of advantages over the use of traditional camera setups. First, the use of reflective tape as background makes the measurement robust. Second, placement of both camera and tape near the floor allows straightforward selection of the events of interest, i.e. walking patterns. Third, the required processing is rather limited, which could allow for energy efficient measurement. There are also some drawbacks of the current setup. For instance, tape color, camera sensor, and algorithm thresholds need to match. For flexible use, we need a higher robustness to variations in lighting. Switching the room lighting on or off can currently stop the application (e.g. soccer) from functioning properly. A remaining issue is camera calibration since the measurement accuracy must be known for applications that measure step size or walking speed. Our simple camera model ignores small rotations. In future work, we wish to investigate the use of a (removable) calibration pattern on the tape and estimate camera rotations in addition to baseline *B* and scale parameter *k*. Another remaining issue is the timing of captured image frames. The cameras in our current setup cannot be triggered for accurate timing of frames. Such cameras exist and should be used for applications that require temporal accuracy.

REFERENCES

1. Project Natal, www.xbox.com/en-US/live/projectnatal/.
2. Zhou, Z., Prugel-Bennett, A., and Damper, R.I. A Bayesian Framework for Extracting Human Gait Using Strong Prior Knowledge. *IEEE. Trans. on Pattern Anal. and Machine Intelligence*, vol.28, no.11, pp. 1738-1752.
3. Cato, T.R. and Zimmerman, T.G. Using Cameras to Monitor Actual Inventory. *US Patent Application 2009/0121017.*
4. OpenCV at Willow Garage: http://www.willowgarage.com/pages/software/opencv
5. Forsyth, D.A. and Ponce, J. *Computer Vision: A Modern Approach*. Prentice Hall, 2003.
6. Begg, R., Best, R., Dell'Oro, L., Taylor, S. Minimum foot clearance during walking: Strategies for the minimization of trip-related falls. *Gait & Posture, 25* (2007), 191-198.