A Survey on Device-Free Passive Localization and Gesture Recognition via Body Wave Reflections

ERIC WENGROWSKI, Rutgers University

Localization of RF signals has been well understood and realized in a multitude of wireless sensor networks. Recently, there has been a surge in research interest surrounding wireless human tracking via RF. This broadly defined domain encompasses areas of research as diverse and expansive as RF Communication Networks, DSP, HCI, and Computer Vision. Traditionally, human tracking is achieved either using a wireless device attached to the subject or utilizing line-of-sight communication. Many existing systems employ wearable wireless tracking devices such as passive tags or active tags that may be monitoring GPS, ultrasonic, or accelerometer data. Obvious drawbacks to wearable tracking devices are the practicalities of users remembering to wear the device, device cost, and power management in active tags. Line-of-sight communication systems usually work in the visible and IR light spectrum and offers relief from these problems, but offer little flexibility around the obvious and inherent line-of-sight constraint. Hence, existing human localization techniques leave much to be desired within the domain of practical home applications. In this article, we explore systems that are able to localize a human in 3D space without a wearable RF device or line-of-sight constraints by measuring body reflections in the non-visible spectrum. Furthermore, we explore how RF-based and near-visible device-free localization can be used for pose estimation, for example in fall detection.

General Terms: RF Communication, Tracking, Localization, 3D Motion Tracking, Wireless Signals, Computer Vision, Infrared, Kinect, Body Reflections, Seeing Through Walls, Pose Recognition, Activity Recognition, Gesture Recognition, Fall Detection

1. INTRODUCTION

Human localization and motion tracking has emerged as a key area of research with a wide array of applications in healthcare monitoring, virtual reality, video games, and social media sharing [Pantelopoulos and Bourbakis 2010]. To a large extent, these applications all fall under the umbrella of Human-Computer Interaction (HCI). However the methods by which human spatial activity is tracked are quite diverse. Traditionally, computer-vision camera systems and mobile RF devices, such as body-worn active RF tags, have been the foremost agents by which real-time human movement is tracked [Moeslund and Granum 2001] [Bulling et al. 2014]. There is a tremendous body of work defining techniques that enable smartphone to track user location and motion. [Herrera et al. 2010] However there has recently emerged a number of techniques by which user location can be estimated without the use of any portable RF device. In these methods, some a signal is broadcasted, and its reflections off the human body are measured. The characteristics of these measured reflections are used to estimate the relative shape and 3D position of the body being tracked. Researchers are essentially treating the user's body as a sort of passive RFID tag. In this way, it is the *user* being tracked rather than an external RF device they are attached to.

Author's addresses: E. Wengrowski, Electrical and Computer Engineering Department, Rutgers University.

DOI: http://dx.doi.org/10.1145/0000000.0000000

In this survey, we explore systems and approaches by which RF signals are used to localize humans without the use of any wearable device. We also explore various approaches by which device-free humans are tracked without traditional stereo computer vision systems. We then compare and contrast these techniques with the methods being developed in the RF domain.

This scope of this survey includes applications such as pose recognition. When a user performs an action, such as sitting down or pointing, they are essentially changing the shape that their body projects. This, in conjunction with locational information, enables many systems to estimate pose as well. [Fadel Adib and Miller 2013] [Fadel Adib and Katabi 2014]

In this article, we will explore the body of research and technologies being developed to enable mobile device-free localization and pose estimation. Broadly defined, these methods can be grouped into 3 categories.

— Line-of-sight sensors operating in the near-visible IR light spectrum such as the Kinect sense the light directly reflected off of users' bodies. Many of these technologies leverage well-established computer vision techniques such as epipolar geometry, but also employ newer, non-traditional methods such as time-of-flight measurements that have recently become more affordable [Moeslund and Granum 2001]. Common challenges with line-of-sight systems include field of view, low light conditions, exterior/interior portability, and privacy concerns. [Winkler and Rinner 2014]

— Radio Tomographic Imaging techniques have been leveraged for human localization and pose detection. [Wilson and Patwari 2011] These systems typically require that an area of interest be outfit with a large network of sensors. This network of N sensors are connected via RF signals forming N^2 links. For each of these links a change in recieved signal strength, called RSS or RSSI, is attributed to human movement. [Wilson and Patwari 2010] [Bocca et al. 2013] [Mager et al. 2013]

— Through-wall RF tracking devices broadcast some signal and measure the reflectance using an array of receivers, much like a tomographic approach. Unlike a tomographic approach, these systems do not require a location to be outfitted with dozens of sensors [Fadel Adib and Miller 2013]. As the name implies, this approach enables out-of-sight measurements to be taken within range. Common challenges with this approach include body part identification (i.e. arm vs. leg), limitations on the number of trackable subjects, and must also include a training set for classification.

High accuracy location and pose estimation is a challenge for all approaches. Since gesture recognition is essentially a classification problem, training sets are used in all approaches.

2. LINE-OF-SIGHT APPROACHES

Line-of-sight approaches to human activity tracking operate in the visible and near-visible spectrum. These approaches leverage computer vision techniques to estimate location, track motion, and characterize gestures. The users' environment is outfitted with one or more cameras. Therefore, the area of sensing is at best limited to the cameras initial field of view. The occurrence of occlusions may limit this field of view further.

In general, computer vision techniques for real-time depth estimation and motion tracking are well defined. [Moeslund and Granum 2001] The areas of interest pertain to feature detection in noisy environments and robust action classification. A large field of research exists related to human identifications in video. [Moeslund et al.

2006] Facial detection has become prevalent [fac] [Yang et al. 2002] [Osuna et al. 1997], as has human detection based on the design and fusion of weak classifiers. [Luke 2010] This body of work has also drastically improved 3D models in the presence of occlusion. [Xia et al. 2011]

We will examine the following vision-based systems and their approaches to robust 3D localization and gesture recognition of a human subject.

2.1. Classification using Human Silhouette

Well-established epipolar geometric techniques make simple 3D scene estimation relatively trivial with vision-based systems. Similarly, tracking the location and motion of a human, as long as they are not obstructed, is also relatively well-established. [Moeslund and Granum 2001]

The projection of the human silhouette can be used in a 2D vision system to reveal human body configurations [shape1] and human motion recognition [Jin and Mokhtarian 2005].

> “Indeed, the human shape will progressively and slowly change during usual activities, while during a fall, it will change drastically and rapidly”. [Rougier et al. 2011]

2.1.1. Robust Video Surveillance for Fall Detection Based on Human Shape Deformation.

The authors of [Rougier et al. 2011] develop a 2D camera system that detect falls during a video sequence by quantifying human shape deformation. The silhouette is obtained using foreground segmentation. Canny edge detection is used to extract edge points from the silhouette.

Moving edge points from two consecutive images are then matched using shape context. [Belongie et al. 2002] The authors employ a matching algorithm using only N points that have minimal matching cost. [Rougier et al. 2011] The Hungarian algorithm has also been suggested for bipartite matching. [Kuhn 1955]

The authors employed two deformation measures to analyze shape change:

- Mean Matching Cost (mentioned earlier)
- Full Procrustes Distance (Proc)

Gaussian mixture modeling is used to classify each activity. The authors examine both the shape deformation as an action occurs as well as post-action movement.

The intended application of the [Rougier et al. 2011] and [Rougier et al. 2007] systems are robust fall detection specifically. However, these techniques could likely be leveraged to recognise poses and actions other than falls.

Using multiple cameras and a voting strategy, the authors achieve an error rate of 4.6% and 3.8% for *mean matching cost* and *full Procrustes distance*, respectively.

2.2. Kinect

Advantages.

The Kinect is a low-cost imaging sensor comprised of software developed at Rare and an infrared (IR) depth camera developed by PrimeSense. [Gill et al. 2011] The Kinect sensor costs approximately $150, significantly less costly than a comparable stereo vision system. It operates in IR on the basis of an actively emitted diffraction grating pattern. This gives the Kinect a massive advantage when sensing the distance of a sparsely textured area such as a blank wall. Also, the IR domain allows the Kinect to operate in low-light conditions without emitting distracting or disruptive visible light. And by operating solely in IR, the Kinect is able to avoid interference from televisions and common household lighting. [Gill et al. 2011]

Disadvantages.

The Kinect sensor has a limited field of view, approximately 60 degrees. The IR sensor experiences significant interference from natural and halogen light, rendering the Kinect essentially useless in the presence of daylight. [Gill et al. 2011] The Kinect sensor has a practical depth range of only 1.2 - 3.5 meters.

2.2.1. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping and Applications.

Although the Kinect is a commercial sensing product, little information about the geometric quality of the data is officially available. However, the accuracy and resolution of the Kinect sensor as a function of depth has been researched. [Khoshelham and Elberink 2012]

> Experimental results show that the random error of depth measurement increases with increasing distance to the sensor, and ranges from a few millimeters up to about 4 cm at the maximum range of the sensor. The quality of the data is also found to be influenced by the low resolution of the depth measurements. [Khoshelham and Elberink 2012]

Comparable geometric investigations and calibrations have been performed on time-of-flight sensors such as the SwissRanger and PMD in previous works. [Kollorz et al. 2008] [Kahlmann and Ingensand 2008] [Kahlmann et al. 2006] [Lichti 2008] [Lindner et al. 2010] [Shahbazi et al. 2011]

The authors of [Khoshelham and Elberink 2012] performed geometric calibration tests on both the Kinect IR camera and a typical RGB camera and found radial lense distortion and decentering to be more severe on the Kinect. Although at a radial distance less than 3̃.6 mm, the Kinect outperformed its RGB counterpart. The authors next compared the point cloud generated by the Kinect sensor with the point cloud generated by a high-end laser scanner. 84% of the point pairs were less than 3 cm apart. Lastly, a Kinectgenerated point cloud was fit to a planar surface of known distance. The results revealed the combined effect of random error and low depth resolution at large distances. The estimated planar surface appeared as several layers of points when seen from a side-view. The authors go on to suggest that fitting geometric object models to the data to reduce the influence of random errors, but suggest that systematic errors can only be resolved through proper calibration procedure. [Khoshelham and Elberink 2012]

2.2.2. Human Detection Using Depth Information by Kinect.

Using depth information alone, the Kinect sensor can be used to detect a human body. Edge information embedded in the depth array is used to locate regions of interest that may indicate the presence of a human. For each image, Canny detection yields an edge map. This is then compared with the Kinects distance map to produce a edge-distance approximation map. 2D chamfer distance matching compares the processed image to a 2D binary head template. Head-like objects are parameterized based on their depth to estimate a 3D head model. The hemispheric models are fitted onto image regions tagged earlier as head-like objects. A threshold is used to determine whether the object is infact a human head, based on squared error between the circular region and the 3D model. A successful match enables us to determine that the entire region shared with this detected human head is likely part of the human body. Whole body contours are extracted from this region to recognise body parts such as hands and feet. The human region can then be tracked across frames. Rather than tracking the human based on color, the motion of the 3D region is estimated using an energy score, which assumes relatively steady changes in motion. The total accuracy of the algorithm using the Kinect sensor was 98.4%. [Xia et al. 2011]

2.2.3. A system for change detection and human recognition in voxel space using the Microsoft Kinect sensor.

There has been tremendous interest in using the Kinect sensor for volumetric scene capture. The authors of [Gill et al. 2011] demonstrate that the Kinect sensor can be used in place of a traditional stereo camera system to accomplish the same fall detection task outlined in [Luke 2010].

2.3. Leap Motion

The Leap Motion controller uses IR light to measure hand position and estimate pose. [Motion 2012] Unlike the Kinect, Leap Motion does not measure the location of, say, a person within a room. Leap is not a full body gesture recognition IR sensor system. However, many authors chose to mention Leap Motion anyway within the context of popular, low-cost, non-wearable, gesture recognition devices. [Fadel Adib and Miller 2013] [Fadel Adib and Katabi 2014] [Meyer]

2.3.1. Accuracy and Robustness of the Leap Motion Controller.

The makers of Leap Motion advertise less than 0.01 mm sensor precision. Empirical studies reveal that Leap Motion is able to measure the position of a static point within 0.2 mm accuracy. For a moving point, Leap Motion is able to localize points in 3D space with an average accuracy of 0.7 mm. [Weichert et al. 2013]

3. RADIO TOMOGRAPHIC IMAGING

Radio Tomography is the process imaging a spatial section or slice using an RF signal. Tomographic measurements are taken from several perspectives to get a more complete view of a scene. In the case of an MRI, a sensor is rotated around the subject. However, when localizing a human subject within the confines of a room or building, usually multiple stationary RF devices are dispersed around the area of interest. These N RF transceivers form a network of N^2 RF signal edges. From [Wilson and Patwari 2011]:

> Indoor radio channel characterization research demonstrates that objects moving near wireless communication links cause variance in RSS measurements [Pratt et al. 2008]. This knowledge has been applied to detect and characterize motion of network nodes and moving objects in the network environment [Woyach et al. 2006]. Polarization techniques have also been used to detect motion [Pratt et al. 2008].

Radio tomographic sensing can been used to detect and localize human movement. First, the sensor network is calibrated. In RSSI tomographic sensing, the received radio strength of each edge is measured and recorded. After the calibration phase, if the received signal strength is significantly decreased, it is attributed to some motion in the area of interest. Leveraging knowledge of sensor and link location, the sensor signal reading is used to estimate movement location. [Wilson and Patwari 2010] [Mager et al. 2013] [Wilson and Patwari 2011]

A major disadvantage of radio tomographic imaging (RTI) is the system size. In each of the systems featured below, the area of coverage must be outfit with dozens of sensors.

Online calibration must be used to for any practical realization of an indoor RTI system due to the highly dynamic nature of indoor environments. [Bocca et al. 2013] The authors of [Wilson and Patwari 2011] [Kaltiokallio et al. 2012] [Zheng and Men 2012] [Edelstein and Rabbat 2013] have all proposed methods to adapt to dynamic environments.

3.1. Human Detection using Radio Tomographic Imaging with Wireless Networks

3.1.1. Experimental Setup.

Experiments were conducted by [Wilson and Patwari 2010] to test the accuracy for which their outdoor system could localize movement from a human body. In these experiments, 28 sensing nodes were uniformly distributed along the perimeter of a 21 x 21 foot area.

> The network comprises TelosB wireless nodes made by Crossbow. Each node operates in the 2.4 GHz frequency band, and uses the IEEE 802.15.4 standard for communication. A base station node listens to all network traffic, then feeds the data to a laptop computer via a USB port for the processing of the images. Since the base station node is within range of all nodes, the latency of measurement retrieval to the laptop is low, on the order of a few milliseconds. If a multihop RTI network were to be deployed, this latency would certainly increase. To avoid network transmission collisions, a simple token passing protocol is used.

The authors of [Wilson and Patwari 2010] used a cylindrical human model where R_H represents the radius of the cylinder.

3.1.2. Reconstruction Results.

For a single image localization of a relatively unobstructed human with $R_H = 1.3$ feet, the squared error was 0.021. The squared error for a comparable two-person single image was 0.036.

It is important to note that the human body not only attenuates a wireless signal, it reflects and scatters it. This has an effect on the RSS at unoccupied sample areas, and is particularly susceptible to error in the presence of interfering objects, such as walls. [Wilson and Patwari 2010]

3.2. Radio Tomographic Imaging for Ambient Assisted Living

The authors of [Bocca et al. 2013] designed a RTI system for single-room indoor use. Self-calibration is designed to make the system more robust to changes in an indoor environment.

3.2.1. Experimental Setup.

Data was collected during a three months deployment of an RTI system composed of 33 nodes in a typical 58 m^2 single floor, one bedroom apartment inhabited by a single person. The experiments were carried out with Texas Instruments CC2531 USB dongle nodes. The nodes are equipped with a low-power, 802.15.4 compliant radio operating in the 2.4 GHz ISM band. The transmit power of the nodes was set to the maximum nominal value, 4.5 dBm. The 802.15.4 standard defines 16 frequency channels, 5 MHz apart and having 2 MHz of bandwidth. The carrier frequency (in MHz) of channel c is: $f_c = 2405 + 5(c11), c[11, 26]$. The nodes are set to communicate on 5 frequency channels, $F = 11, 15, 18, 21, 26$. The nodes employed a multi-spin communication protocol outlined in [Bocca et al. 2013].

3.2.2. Results.

The average localization error is 0.23 m. It has to be noted that the localization error remains below 0.40 m in 13 of the 14 points of evaluation. The largest error, 0.92 m, is measured when the human subject stands near a large marble counter, which has a remarkable impact on the propagation of the radio signals. [Bocca et al. 2013]

3.3. Fall Detection Using RF Sensor Networks

RTI networks measuring RSS can be used to classify rough pose and detect falls. By placing sensors at multiple heights, a subjects telemetry can be estimated. However, the obvious drawback for this approach is the need for at least twice as many sensors.

3.3.1. Dual-Height Sensor Network Layout.

In [Mager et al. 2013], the authors designed a dual-height RTI sensor network. In their testbed, sensors were placed at two heights at each (x, y) location. The first layer of sensors were placed 17 cm above the floor. The second layer were placed directly above, at a height of 140 cm.

In a dual-layer RTI system, RSS is separately determined for each of the two sensor layers, ignoring the crossover links spanning multiple heights. However, the diagonal RF crosslink between the two sensor heights provides additional information about a persons vertical position. This approach is known as 3D RTI. In the approach adopted by [Mager et al. 2013], five-layers of RTI estimates were employed when using RTI in 3D for fall detection.

3.3.2. Pose State Machine.

A 3-state classification system was used to recognise pose. The three states were standing, mid-level, and lying down. Since the system measures RSS at 5 height layers, the mid-level pose state allowed the system to classify actions such as sitting down and lying on a bed without thrashing between standing and lying down. A hidden markov method was used to model the transition from the standing state to the falling state. A fall is determined by the transition time from the standing state to the lying-down state. [Mager et al. 2013]

3.4. See-through walls: Motion Tracking Using Variance-Based Radio Tomography Networks

When a stationary object obstructs the link in a through-wall environment, the change in mean RSS is unpredictable. When an object moves, the variance of the obstructed links RSS provides a more reliable metric.

3.4.1. VTRI - Variance-Based Radio Tomographic Imaging.

Rather than relying solely on the magnitude of Received Signal Strength (RSS) in decibels, the goal of a VRTI system is to use a vector s of RSS variance measurements on M links in a wireless network to determine an image vector x that describes the presence of motion occurring within N voxels of a physical space. When a moving object affects the amplitude or phase of one or more multipath components over time, the phasor sum of all multipath at the receiver experiences changes, and higher RSS variance is observed. The amount of RSS variance relates to the physical location of motion, and an image representing motion is estimated using measurements from many links in the wireless network. [Wilson and Patwari 2011]

3.4.2. Experimental Configuration.

A network of 34 sensor nodes was nonuniformly distributed around the perimeter of a 24 x 18 ft room. Each node was placed behind a brick wall. The sensor hardware is congruent to those used in the [Bocca et al. 2013] experiment.

3.4.3. Results.

The average tracking error in the overhead-view x, y coordinate plane is 2.07 feet. The average localization error of an object moving in-place is 1.46 feet. [Wilson and Patwari 2011]

3.4.4. Concluding Remarks

> Tracking multiple moving targets through dense walls is a challenging and open topic for future research. When multiple people move within a surveillance area, the accuracy of a VRTI image is dependent on the separation of the targets. Additionally, when multiple walls or walls constructed with dense materials surround the surveillance area, the amount of power that radiates into the area is reduced. [Wilson and Patwari 2011]

VRTI techniques do allow human movement to be tracked through walls. Since this system does not measure 3D location, and due to the low localization accuracy, it is unlikely that pose recognition such as fall detection could be performed. Additionally, this system falls short with the inherent drawbacks of all tomographic approaches: it requires large network of many sensors, and an extensive, location-dependent, training period is needed. However, being that this body of work does allow through-wall localization, it closely links RF tomography with the research and methods explored in the next section.

4. THROUGH-WALL RF TRACKING

Through-wall RF tracking devices broadcast an RF signal and measure its reflectance, similarly to tomographic approaches. Unlike a tomographic approach, these systems are smaller, and do not require a broad dispersal of sensors around an area of interest, similarly to visual and near-visual line-of-sight approaches. However, as the name implies, through-wall RF tracking devices are able to make out-of-sight measurements, unlike a camera-based approach. All three of these approaches require training to accurately classify pose. Through-wall RF tracking devices require initial calibration of the area of observance. Common challenges with these approaches include body part identification (i.e. arm vs. leg) and limitations on the number of trackable subjects.

4.1. WiTrack

WiTrack is a through-wall RF tracking system that tracks the 3D position of a single moving body in real-time. WiTrack operates by measuring the directional wireless signal reflected off of the users body at the 5.56 GHz to 7.25 GHz range. WiTrack is able to perform non-line-of-sight tracking. More specifically, it is able to measure 3D localization through a household wall. WiTrack also performs coarse estimation of pose, such as the direction of a pointed arm, or whether the user is lying down. [Fadel Adib and Miller 2013]

4.1.1. Wireless Hardware.

WiTrack wireless hardware components consist of four directional antennas. One antenna is used for transmission, and the other three are used for reception. The arrangement of each forward facing directional antenna is in a T shape. The transmit antenna is placed in the center. Two receiving antennas are each placed 5 cm to the right and 5 cm to the left of the transmitter. The final receiving antenna is placed 5 cm below the transmit antenna. These wireless devices collectively sweep a total bandwidth of 1.69 GHz from 5.56 GHz to 7.25 GHz, and transmit at 0.75 milliwatt. A reference signal sweeps from 136.5181.25 MHz to generate an FMCW signal that sweeps from 5.467.25 GHz. This FMCW signal is transmitted over the air using WA5VJB directional antennas. [Fadel Adib and Miller 2013]

4.1.2. Time of Flight.

WiTrack makes time-of-flight (TOF) measurements to estimate distance. FMCW (frequency modulated carrier wave) signals and techniques facilitate this. FMCW transmits a narrowband signal (e.g., a few KHz) whose carrier frequency changes linearly with time. After the signal is transmitted and a portion of the it is reflected back toward the receiver, the FMCW maps differences in time to shifts in the carrier frequency. These frequency shifts are easy to measure in radio systems by looking at the spectrum of the received signal. In this way, differences in reflection time can simply be measured as shifts in carrier frequency. With reflection time known, the distance of reflecting objects can be calculated trivially using the speed of light.

$$\text{round trip distance} = c \times TOF = c \times \frac{\Delta f}{slope}$$

4.1.3. Multipath Effects.

Multipath effects come in two flavors: static and dynamic.

With static multipath, the goal is to distinguish RF signals reflected off humans versus other objects in the environment like walls and furniture. Otherwise, reflections off of obstacles would mask the signal coming from the human. This phenomenon is known as Flash Effect. To control for this, WiTrack operates under the assumption that: there is only 1 human in an environment, that human moves, and that human is responsible for all movement. Following that logic, WiTrack ignores any signal that does not change over time.

> Hence, we eliminate the power from these static reectors by simply subtracting the output of the FFT in a given sweep from the FFT of the signal in the previous sweep. This process is called background subtraction because it eliminates all the static reectors in the background.

Dynamic multipath effects occur when the signal reflected off of a human also reflects off other objects in the environment before reaching the WiTrack antennas. It is possible that the signal reflected off of another object is stronger than the signal directly linking WiTrack and the moving human. In order to localize, WiTrack must be able to accurately estimate a humans direction and radial distance. So identifying the most direct path between the WiTrack antennas and the human subject is key. To resolve this problem, WiTrack leverages the TOF characteristics of its FMCW. The underlying assumption is signal directly reflected off the human will have the smallest frequency shift after background subtraction.

4.1.4. Robustly Estimating Location via Interpolation and Filtering.

Since WiTrack can only sense moving objects, it would quickly lose track of a subject as soon as it stops moving. To combat this, WiTrack stores the subjects last known location. If no moving signal is present, WiTrack interpolates based on that last known location. Large, rapid jumps in locationional estimation are disregarded as noise. Kalman filtering is used to smooth the location estimates. [Fadel Adib and Miller 2013]

4.1.5. Localization Accuracy.

A $6x5m^2$ test area was used. For all points in testing, the human subjects remained between a 3 meter and 9 meter range from WiTrack.

The VICON T Series motion capture system was used to measure ground truth. VICON performs 3D localization measurements with sub-centimeter accuracy [Peak 2005], and was used to locate the center of the body.

For line-of-sight measurements, WiTrack localizes the center of the body to within a median of 9.9 and 8.6 cm in the x and y dimensions, and 17.7 cm in the z dimension. For through-wall experiments, WiTrack localizes the center of the body to within a median of 13.1, 10.25 cm, and 21.0 cm in the x, y, and z dimensions, respectively. [Fadel Adib and Miller 2013]

As an important note, the authors define the plane that WiTracks antennas are all along as the *x*-axis. The standing height of a human is largest in the *z*-dimension.

> The median accuracy changes by 5 to 10 cm for distances that are 3 to 11 m away from the device. As expected, the further the human moves from the device, the larger the estimation error. This increase in error with distance is expected since as the distance gets larger the signal gets more attenuated. However, a second reason stems from the geometry of the ellipsoid-based localization model. Given the equations of the ellipsoid, the TOF multiplied by the speed of light is equal to the major axis of the ellipsoid/ellipse that describes the users location, and the antenna separation is the distance between the foci. For a fixed antenna separation, as the distance/TOF increases the ellipsoids surface increases, increasing the overall space of potential locations. [Fadel Adib and Miller 2013]

4.1.6. Pose Recognition.

WiTrack is designed to recognise pose. The two pose detection applications are fall detection and pointing direction.

WiTrack is only able to estimate pointing direction if only the users arm is the sole moving section of their body. To segment an arm, the authors leverage the assumption that the area of an arm is significantly smaller than the area of the entire body. Additionally, users are instructed to pause for 1 second between raising and lowering an arm to clearly isolate each movement. The location estimates of the moving hand are regressed to determine hand position and pointing direction. Empirical analysis shows that the median orientation error is 11.2 degrees, and the 90th percentile error is 37.9 degrees. [Fadel Adib and Miller 2013]

The authors of WiTrack define a fall to be a rapid change in *z*-axis elevation where the final resting place is close to the ground. To experimentally test the accuracy of WiTracks fall detection algorithm, participants were asked to perform four distinct activities: walk, sit on a chair, sit on the floor, and simulate a fall. The authors perform 132 experiments in total, 33 for each activity. Out of the 32 detected falls, only 31 were true falls. Additionally, out of 33 true falls, 31 were detected. Thus, the precision of the fall detection algorithm is 96.9

4.1.7. Concluding Remarks.

WiTrack uses FMCW RF signals to perform time-of-flight estimation. WiTrack is able to perform through-wall 3D localization of a single person in a windowless room with a median error of 10 to 13 cm in the x and y dimensions, and 21 cm in the z dimension. The system makes its measurements using 4 antennas, all fixed to a single T-shaped structure. The system does not require environment-dependent calibration. The device has no way of distinguishing different body parts (i.e. arm vs. leg). WiTrack can only sense a subject when that subject is moving. WiTracks accuracy and reliability for pose recognition are discussed in the previous subsection. The transmitting power of WiTrack is 0.75 milliwatts. [Fadel Adib and Miller 2013]

4.2. WiZ

The WiZ RF localization system is largely an extension of WiTrack. Similarly to WiTrack, WiZ is able to deliver through-wall motion tracking and gesture recognition. Both WiTrack and WiZ use time-of-flight measurements enabled by a FMCW. And both systems use a planar array of antennas that operate in the same frequency band (5.46-7.25 GHz), and transmit at 0.75 mW. However, there are some stark differences between WiTrack and . Unlike WiTrack, WiZ does not localize users in 3D space, only birds-eye 2D space. Because users z-axis position is not tracked, WiZ is unable to perform fall detection. Instead, WiZ is able to simultaneously track 3 to 5 different humans, whereas WiTrack is only able to track 1 user. WiZ is also able to localize static humans based on their breathing movement. Both WiZ and WiTrack use VICON [Peak 2005] to establish ground-truth location. [Fadel Adib and Miller 2013] [Fadel Adib and Katabi 2014]

4.2.1. Wireless Hardware.

consists of a FMCW radio, USRP N210 software radios with LFRX-LF daughterboards, and directional antennas. The FMCW radio generates a signal that sweeps 5.46 - 7.25 GHz every 2.5 milliseconds. WiZ is a multi-antenna system. It has ve transmit antennas (Tx) and ve receive antennas (Rx). These antennas are directional; each of them is $3cm \times 3.4cm$. They are all stacked on a single planar platform that measures $2 \times 1m$. Each antenna transmits less than 0.75 milliWatts of power. [Fadel Adib and Katabi 2014]

4.2.2. Multi-Shift FMCW.

WiZ incorporates a technique dubbed multi-shift FMCW, a multi-antenna extension to FMCW (discussed previously with WiTrack) where the signal transmitted by different antennas is structured in a particular way to differentiate people and eliminate the impact of ghost TOF (time-of-flight) readings that do not correspond to physical targets. As the number of users in a sensing area increases, TOF measurements from a larger number of Tx-Rx pairs are needed to localize them. In the current implementation, WiZ has 25 Rx-Tx pairs (5 Rx × 5 Tx). Furthermore, WiZ receiver antennas must be able to distinguish and identify the signal sensed by each transmitting antenna.

The authors of [Fadel Adib and Katabi 2014] address this challenge with multi-shift FMCW. As mentioned above, FMCW consists of a continuous linear frequency sweep. The round-trip TOF is calculated when the FMCW hits a body and get reflected back to Rx. The authors make assumptions about the maximum TOF that WiZ expects to encounter in a typical indoor environment, τ. Each transmitter is frequency shifted by τ so that the signal from Tx1 would be TOF_1, and the signal from Tx2 would be $TOF_2\prime = TOF_2 + \tau$. [Fadel Adib and Katabi 2014]

4.2.3. SSC - Successive Silhouette Cancellation.

Just like WiTrack, WiZ is faced with the challenge of resolving dynamic multipath. [Fadel Adib and Miller 2013] But since WiZ aims to localize many users, it must also confront the near-far problem. The near-far problem is a phenomenon where a person who is much closer to the antennas will have much stronger reflections than someone who is farther away. This may results in the reflections of the far person being masked by the reflections of the close person. WiZ employs a technique known as Successive Silhouette Cancellation (SSC) to address this.

> To deal with this near-far problem, rather than localizing all the people in one shot, WiZ performs Successive Silhouette Cancellation (SSC). SSC is inspired by Successive Interference Cancellation whereby the receiver decodes the signal with the highest SNR, then re-encodes it and subtracts it out from the received signal, and proceeds to decode the signal with the second-highest SNR, then repeats the same procedure until it has decoded all interferers. The main difference is that decoding in our context means localizing the person using her TOF.

s SSC algorithm consists of 4 phases:

(1) SSC Detection: This involves finding the location of the strongest user FMCW reflections by overlaying the heatmaps of all Tx-Rx pairs.
(2) SSC Re-mapping: A persons location estimates, d_min and d_max, are mapped to the set of TOFs, TOF_min and TOF_max, that would have generated that location at each transmit-receive pair.
(3) SSC Cancellation: This involves subtracting the impact of the persons signal from the TOF profiles of all TX-Rx pairs.
(4) Iteration: After cancellation, the obtained TOF profiles are used to re-compute the heatmaps, overlay them, and repeat the process of finding the location of the next strongest reflector. . . .

The reflections of more than one person may accidentally be dismissed in the cancellation phase. The extra person is more likely to be removed if they are close to the intended target. It should also be noted that the heatmap generated is much noisier after each subsequent iteration because the distant persons FMCW signals are weaker. [Fadel Adib and Katabi 2014]

During the iteration phase, the SNR (signal-to-noise ratio) is compared to a threshold to determine whether or not an additional person is present or not.

applies a Kalman filter and perform outlier rejection to reject impractical location jumps. For moving humans, the motion path is analysed to associate distinct users with their own trajectory after crossing paths. [Fadel Adib and Katabi 2014]

is able to recognise pointing gestures using nearly identical methods employed in WiTrack [Fadel Adib and Miller 2013], which are mentioned in the previous section. However, WiZ distinguishes between each users silhouette, and is therefore able to recognise different users pointing gestures simultaneously.

4.2.4. Static Localization Based on Breathing.

Unlike WiTrack, WiZ is able to localize static users. WiTrack resolves static multipath by performing a background subtraction. [Fadel Adib and Miller 2013] However this method also eliminates the signal reflected off of non-moving users. WiZ resolves this problem by sampling with 2 differently sized time windows. The first time window is 12.5 milliseconds, which is used to accurately capture human movement such as walking or pointing. The second time window is 2.5 seconds, which is used to capture much small human movement such as breathing. WiZ uses this minute movement to identify and localize users who are standing still or sitting. This allows WiZ to count the number of breaths that a user takes as well.

4.2.5. Performance Evaluation.

Vicon [Peak 2005] was used to establish ground truth locations for performance evaluations. WiZ is able to simultaneously track the motion of 4 and 3 unique users with line-of-sight and through-wall approaches, respectively. For static users, WiZ uses breathing-based localization techniques to localize 5 and 4 unique users using respective line-of-sight and through-wall approaches. The authors of WiZ define reliable detection as successfully tracking the correct number of people with a probability of 0.98 or higher.

For same-room measurements, the median location error is 8.5 cm in x dimension and 6.4 cm in y dimension WiZ for the rst user detected, and decreases to 15.9 cm in x and 7.2 cm in y for the last detected user. For through-wall measurements, the median location error is 8.4 cm and 7.1 cm in x/y for the rst user detected, and decreases to 16.1 cm and 10.5 cm in x/y for the last detected user. The accuracy in the y dimension is better than the accuracy in the x dimension because WiZs antennas are all arranged along the y = 0 axis.

For both line-of-sight and through-wall, WiZs breathing-based localization accuracy has a median of 7.24 cm and 6.3 cm in x/y for the nearest person and 18.31 cm to 10.85 cm in x/y for the furthest person. WiZ correctly counted the number of breaths for over 97

For pointer recognition, the 3D pointing direction was decomposed into two angles: θ and ψ. θ is the projection of the pointing direction on the x y plane, and ψ is the pointing direction in the r z plane. The median orientation error in θ goes from 8.2 degrees to 12.4 degrees from the rst to the third person, and from 12 degrees to 16 degrees in ϕ. [Fadel Adib and Katabi 2014]

4.2.6. Concluding Remarks.

WiZ is able to localize multiple moving and nonmoving humans through-wall with RF body reflections with centimeter accuracy. Although it is authored by the same group, WiZ distinguishes itself from WiTrack with support for multiple users and static localization. However, WiZ does not track user height in 3D space, and therefore does not support the elevation-based fall detection algorithm employed by WiTrack. [Fadel Adib and Miller 2013] Similarly to WiTrack, WiZ is unable to differentiate body parts (i.e. arm vs. leg). Furthermore, WiZ can not accurately track more than 5 users at once. Finally, WiZs antenna array measures $2 \times 1m$, which may be too cumbersome for practical use in some environments.

5. OVERALL PERFORMANCE EVALUATION

These accuracy of each of the systems examined in this survey are significant. However it is nearly impossible to directly compare each of these systems to each other based on that accuracy metric alone. For instance, each of these systems has been tested in totally different environments, each with unique size boundaries, obstacles, construction materials, and amount of external noise. More importantly, these systems are not all measuring the same thing. For Example, WiTrack [Fadel Adib and Miller 2013] measures accuracy of localization and fall detection classification, but the authors of [Rougier et al. 2011] measure the accuracy of fall detection classification only. When evaluating each of these systems, it is important to understand the context by which they were designed to operate in. And more importantly, the general trade-offs and constraints of each methodology.

6. CONCLUSION

In this survey, we explore various topical research systems and methods by which humans are tracked without any wireless device. The accuracies, practicalities, and trade-offs of various approaches are explored and compared. The tracking methods are sorted into one of three categories: near-visual line-of-sight, radio tomographic, and non-tomographic through-wall sensing. Within each category, research methods are explained and analyzed. Each category also includes analysis of at least one pose recognition application, such as fall detection.

REFERENCES

(????).

S. Belongie, J. Malik, and J. Puzicha. 2002. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24, 4 (Apr 2002), 509–522. DOI:http://dx.doi.org/10.1109/34.993558

Maurizio Bocca, Ossi Kaltiokallio, and Neal Patwari. 2013. Radio Tomographic Imaging for Ambient Assisted Living. In *Evaluating AAL Systems Through Competitive Benchmarking*, Stefano Chessa and Stefan Knauth (Eds.). Communications in Computer and Information Science, Vol. 362. Springer Berlin Heidelberg, 108–130. DOI:http://dx.doi.org/10.1007/978-3-642-37419-7_9

Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Comput. Surv.* 46, 3, Article 33 (Jan. 2014), 33 pages. DOI:http://dx.doi.org/10.1145/2499621

Andrea Edelstein and Michael Rabbat. 2013. Background subtraction for online calibration of baseline rss in rf sensing networks. *Mobile Computing, IEEE Transactions on* 12, 12 (2013), 2386–2398.

Dina Katabi Fadel Adib, Zachary Kabelac and Robert C. Miller. 2013. *3D Tracking via Body Radio Reections*. MIT CSAIL Technical Report TR-2013-030. Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/82913

Zachary Kabelac Fadel Adib and Dina Katabi. 2014. *Multi-Person Motion Tracking via RF Body Reflections*. MIT CSAIL Technical Report MIT-CSAIL-TR-2014-008. Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/86299

T. Gill, J.M. Keller, D.T. Anderson, and R.H. Luke. 2011. A system for change detection and human recognition in voxel space using the Microsoft Kinect sensor. In *Applied Imagery Pattern Recognition Workshop (AIPR), 2011 IEEE*. 1–8. DOI:http://dx.doi.org/10.1109/AIPR.2011.6176347

Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The¡ i¿ Mobile Century¡/i¿ field experiment. *Transportation Research Part C: Emerging Technologies* 18, 4 (2010), 568–583.

Ning Jin and F. Mokhtarian. 2005. Human motion recognition based on statistical shape analysis. In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*. 4–9. DOI:http://dx.doi.org/10.1109/AVSS.2005.1577234

Timo Kahlmann and Hilmar Ingensand. 2008. Calibration and development for increased accuracy of 3D range imaging cameras. *Journal of Applied Geodesy* 2, 1 (2008), 1–11.

Timo Kahlmann, Fabio Remondino, and H Ingensand. 2006. Calibration for increased accuracy of the range imaging camera swissrangertm. *Image Engineering and Vision Metrology (IEVM)* 36, 3 (2006), 136–141.

O. Kaltiokallio, M. Bocca, and N. Patwari. 2012. Follow @grandma: Long-term device-free localization for residential monitoring. In *Local Computer Networks Workshops (LCN Workshops), 2012 IEEE 37th Conference on*. 991–998. DOI:http://dx.doi.org/10.1109/LCNW.2012.6424092

Kourosh Khoshelham and Sander Oude Elberink. 2012. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors* 12, 2 (2012), 1437–1454. DOI:http://dx.doi.org/10.3390/s120201437

Eva Kollorz, Jochen Penne, Joachim Hornegger, and Alexander Barke. 2008. Gesture recognition with a time-of-flight camera. *International Journal of Intelligent Systems Technologies and Applications* 5, 3 (2008), 334–343.

H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97. DOI:http://dx.doi.org/10.1002/nav.3800020109

Derek D Lichti. 2008. Self-calibration of a 3D range camera. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 37 (2008), 927–932.

Marvin Lindner, Ingo Schiller, Andreas Kolb, and Reinhard Koch. 2010. Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding* 114, 12 (2010), 1318–1328.

Robert H Luke. 2010. *A system for change detection and human recognition in voxel space using stereo vision*. Ph.D. Dissertation. University of Missouri–Columbia.

Brad Mager, Neal Patwari, and Maurizio Bocca. 2013. Fall detection using RF sensor networks. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*. IEEE, 3472–3476.

Roland Meyer. The Use of Wireless Signals for Sensing and Interaction. (????).

Thomas B Moeslund and Erik Granum. 2001. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding* 81, 3 (2001), 231–268.

Thomas B. Moeslund, Adrian Hilton, and Volker Krger. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104, 23 (2006), 90 – 126. DOI:http://dx.doi.org/10.1016/j.cviu.2006.08.002 Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.

Leap Motion. 2012. Leap. *URL: https://www. leapmotion. com/[last accessed 2013-02-04]* (2012).

E. Osuna, R. Freund, and F. Girosi. 1997. Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. 130–136. DOI:http://dx.doi.org/10.1109/CVPR.1997.609310

Alexandros Pantelopoulos and Nikolaos G. Bourbakis. 2010. A Survey on Wearable Sensor-based Systems for Health Monitoring and Prognosis. *Trans. Sys. Man Cyber Part C* 40, 1 (Jan. 2010), 1–12. DOI:http://dx.doi.org/10.1109/TSMCC.2009.2032660

Vicon Peak. 2005. Vicon Motion Capture System. (2005).

T Pratt, S Nguyen, and BT Walkenhorst. 2008. Dual-polarized architectures for sensing with wireless communications signals. In *Military Communications Conference, 2008. MILCOM 2008. IEEE*. IEEE, 1–6.

C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. 2007. Fall Detection from Human Shape and Motion History Using Video Surveillance. In *Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on*, Vol. 2. 875–880. DOI:http://dx.doi.org/10.1109/AINAW.2007.181

C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. 2011. Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. *Circuits and Systems for Video Technology, IEEE Transactions on* 21, 5 (May 2011), 611–622. DOI:http://dx.doi.org/10.1109/TCSVT.2011.2129370

Mozhdeh Shahbazi, Saeid Homayouni, Mohammad Saadatseresht, and Mehran Sattari. 2011. Range camera self-calibration based on integrated bundle adjustment via joint setup with a 2D digital camera. *Sensors* 11, 9 (2011), 8721–8740.

Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Denis Fisseler. 2013. Analysis of the accuracy and robustness of the leap motion controller. *Sensors (Basel, Switzerland)* 13, 5 (2013), 6380.

Joey Wilson and Neal Patwari. 2010. Radio tomographic imaging with wireless networks. *Mobile Computing, IEEE Transactions on* 9, 5 (2010), 621–632.

Joey Wilson and Neal Patwari. 2011. See-through walls: Motion tracking using variance-based radio tomography networks. *Mobile Computing, IEEE Transactions on* 10, 5 (2011), 612–621.

Thomas Winkler and Bernhard Rinner. 2014. Security and Privacy Protection in Visual Sensor Networks: A Survey. *ACM Comput. Surv.* 47, 1, Article 2 (May 2014), 42 pages. DOI:http://dx.doi.org/10.1145/2545883

Kristen Woyach, Daniele Puccinelli, and Martin Haenggi. 2006. Sensorless sensing in wireless networks: Implementation and measurements. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on*. IEEE, 1–8.

Lu Xia, Chia-Chih Chen, and J.K. Aggarwal. 2011. Human detection using depth information by Kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. 15–22. DOI:http://dx.doi.org/10.1109/CVPRW.2011.5981811

Ming-Hsuan Yang, D. Kriegman, and N. Ahuja. 2002. Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24, 1 (Jan 2002), 34–58. DOI:http://dx.doi.org/10.1109/34.982883

Yi Zheng and Aidong Men. 2012. Through-wall tracking with radio tomography networks using foreground detection. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*. IEEE, 3278–3283.

ACKNOWLEDGMENTS

The author would like to thank Dr. Yanyong Zhang of Rutgers University.