Practical Motion Capture in Everyday Surroundings

Daniel Vlasic Rolf Adelsberger[†‡] Giovanni Vannucci[†] John Barnwell[†] Markus Gross[‡]
Wojciech Matusik[†] Jovan Popović

Computer Science and Artificial Intelligence Laboratory [†]Mitsubishi Electric Research Laboratories [‡]ETH Zürich
Massachusetts Institute of Technology

Figure 1: Traditional motion-capture systems excel at recording motions within lab-like environments but struggle with recording outdoor activities such as skiing, biking, and driving. This limitation led us to design a wearable motion-capture system that records human activity in both indoor and outdoor environments.

Abstract

Commercial motion-capture systems produce excellent in-studio reconstructions, but offer no comparable solution for acquisition in everyday environments. We present a system for acquiring motions almost anywhere. This wearable system gathers ultrasonic time-of-flight and inertial measurements with a set of inexpensive miniature sensors worn on the garment. After recording, the information is combined using an Extended Kalman Filter to reconstruct joint configurations of a body. Experimental results show that even motions that are traditionally difficult to acquire are recorded with ease within their natural settings. Although our prototype does not reliably recover the global transformation, we show that the resulting motions are visually similar to the original ones, and that the combined acoustic and inertial system reduces the drift commonly observed in purely inertial systems. Our final results suggest that this system could become a versatile input device for a variety of augmented-reality applications.

CR Categories: I.3.7 [Computer Graphics]: Three Dimensional Graphics and Realism—Animation

Keywords: Augmented Reality, Motion Capture

1 Introduction

Motion data has revolutionized computer animation in the last decade. Techniques that edit, transform, interpolate, and recompose motion data can now generate novel animations of unprecedented quality. Their success is readily apparent in films such as Polar Express and Lord of the Rings, which transfer expressive performances of real actors onto digital characters both real and fantastic. An entire industry has emerged in support of these activities, and numerous recordings of human performances are now available in large motion repositories (e.g., *mocap.cs.cmu.edu* and *www.moves.com*).

However, the majority of current acquisition systems inhibit broader use of motion analysis by requiring data collection within restrictive lab-like environments. As a result, motions such as skiing and driving are simply never acquired, while others like cycling and playing football are not recorded in their natural competitive setting. Furthermore, recording the activities, routines, and motions of a human for an entire day is still challenging.

In this paper, we explore the design of a wearable self-contained system that is capable of recording and reconstructing everyday activities such as walking, biking, and exercising. Our design minimizes discomfort and maximizes recording time by prioritizing light-weight components with low power requirements. It records acoustic and inertial information from sensors worn on the body. Inertial measurements are provided by miniature gyroscopes and accelerometers no more than a few millimeters in size. Each sensor also includes a miniature microphone, which is used to record distances between pairs of sensors on the body. These distance measurements reduce the drift common to purely inertial systems.

The reconstruction algorithm estimates body postures by combining inertial and distance measurements with an Extended Kalman Filter that incorporates the body's joint structure. Although it lacks the information to recover global translation and rotation, our approach reconstructs sequences of full body postures that are visually similar to the original motions.

Our system is not the first acoustic-inertial tracker, but it is the first such system capable of reconstructing configurations for the entire body. We show that these reconstructions are most accurate when combining information from all three sensor types: gyroscopes, accelerometers, and distance measurements. The best reconstructions are still not perfect, but their quality, along with the small size and improved versatility, suggest that our system may lead to new applications in augmented reality, human-computer interaction, and other fields.

2 Previous Work

Several motion capture technologies have been proposed in the last two decades. The advantages and disadvantages of the dominant approaches are argued in several excellent surveys [Meyer et al. 1992; Frey 1996; Hightower and Borriello 2001; Welch and Foxlin 2002]. In this brief summary, we review optical, image-based, mechanical, magnetic, inertial, acoustic, and hybrid systems, mentioning a few exemplary systems in each category.

Optical motion capture systems [Woltring 1974; Bishop 1984] and modern systems manufactured by Vicon (*vicon.com*), Codamotion (*codamotion.com*), and PhaseSpace (*phasespace.com*), track retroreflective markers or light-emitting diodes placed on the body. Exact 3D marker locations are computed from the images recorded by the surrounding cameras using triangulation methods. These systems are favored in the computer-animation community and the film industry because of their exceptional accuracy and extremely fast update rates. The major disadvantages of this approach are extreme cost and lack of portability. To reduce cost and improve portability, some systems use a small number of markers in conjunction with standard video cameras. For example, Yokokohji and colleagues [2005] capture arm motions with a head-mounted camera.

Image-based systems [Bregler and Malik 1998; Davison et al. 2001; Chen et al. 2005] use computer vision techniques to obtain motion parameters directly from video footage without using special markers. These approaches are less accurate than optical systems, however, they are more affordable and more portable. Still, they are not entirely self-contained since they rely on one or more external cameras. Furthermore, they suffer from line-of-sight problems, especially in the case of monocular video.

Mechanical systems, such as Meta Motion's Gypsy™ (*metamotion.com*), require performers to wear exoskeletons. These systems measure joint angles directly (e.g., using electric resistance), rather than estimating the positions of points on the body, and can record motions almost anywhere. Exoskeletons are uncomfortable to wear for extended time periods and impede motion, although these problems are alleviated in some of the modern systems, such as Measurand's ShapeWrap™ (*measurand.com*).

Magnetic systems, such as MotionStar® by Ascension Technology Corporation (*ascension-tech.com*), detect the position and orientation using a magnetic field (either the Earth's magnetic field or the field generated by a large coil). These systems offer good accuracy and medium update rates with no line-of-sight problems. However, they are expensive, have high power consumption, and are sensitive to the presence of metallic objects in the environment.

Inertial motion capture systems, such as Xsens's Moven (*xsens.com*) and Verhaert's ALERT system (*verhaert.com*), measure rotation of the joint angles using gyroscopes or accelerometers placed on each body limb [Miller et al. 2004]. Like the mechanical systems, they are portable, but cannot measure positions and distances directly for applications that must sample the geometry of objects in the environment. More importantly, the measurements drift by significant amounts over extended time periods. In addition, the motion of the root cannot be reliably recovered from inertial sensors alone, although in some cases this problem can be alleviated by detecting foot plants [Foxlin 2005].

Acoustic systems use the time-of-flight of an audio signal to compute the marker locations. Most current systems are not portable and handle only a small number of markers. With the Bat system [Ward et al. 1997], an ultrasonic pulse emitter is worn by a user, while multiple receivers are placed at fixed locations in the environment. A system by Hazas and Ward [2002] extends ultrasonic capabilities by using broadband signals; Vallidis [2002] alleviates occlusion problems with a spread-spectrum approach; Olson and colleagues [2006] are able to track receivers without known emitter locations. The Cricket location system [Priyantha et al. 2000] fills the environment with a number of ultrasonic beacons that send pulses along with RF signals at random times in order to minimize possible signal interference. This allows multiple receivers to be localized independently. A similar system is presented by Randell and Muller [2001], in which the beacons emit pulses in succession using a central controller. Lastly, the WearTrack system [Foxlin and Harrington 2000], developed for augmented reality applications, uses one ultrasonic beacon placed on the user's finger and three fixed detectors placed on the head-mounted display. This system can track the location of the finger with respect to the display, based on time-of-flight measurements.

Hybrid systems combine multiple sensor types to alleviate their individual shortcomings. They aim to improve performance, rather than decrease cost and increase portability. For example, an acoustic-inertial system, Constellation™, has been developed for indoor tracking applications [Foxlin et al. 1998]. The system corrects inertial drift using ultrasonic time-of-flight measurements to compute exact distances between receivers and ultrasonic beacons placed at known locations. Another acoustic-inertial system [Ward et al. 2005] uses a wrist-worn microphone and a 3-axis accelerometer for gesture recognition. Similarly, MERG sensors [Bachmann 2000] enable inertial-magnetic systems that account for the drift by using a reference magnetic field. In the same manner, Hy-BIRD™ by Ascension Technology Corporation (*ascension-tech.com*) combines optical and inertial technologies to tackle occlusion problems. Finally, a combination of image-based and inertial tracking is used for sign language recognition [Brashear et al. 2003].

3 System Prototype

Our wearable motion-capture system consists of ultrasonic and inertial subsystems, a driver box that controls their operation, and a storage device that records the data. During operation, the data from the two independent subsystems (the ultrasonic subsystem used for distance measurements and the inertial subsystem used for measurements of accelerations and rotation rates) are acquired at each sensor board, encoded, and jointly transmitted to the driver box. The driver box samples the signals from all the sensor boards and transfers them onto the storage drive for off-line signal processing and pose estimation. This section describes our hardware and its operation at a high level. More details, including individual part numbers, can be found in the document by Adelsberger [2007].

3.1 System Components

First, we outline the architecture of our system and describe the individual components. As illustrated in Figure 2, our system is controlled by a custom-built driver box connected to a laptop using a USB interface. The driver box is also connected to eight ultrasonic sources and eighteen sensor boards using shielded 3-wire cables. All sensors are attached to the user's garment. The driver box provides pulse signals to the ultrasonic sources, polls data from the sensors, and provides power to the sensor boards. The driver box is powered by a rechargeable Lithium-Ion battery pack with 4.4 AHr capacity, which provides several hours of safe operation (the current drawn by our system is 1.5A). A laptop records the data on a hard drive but is not used for any processing. We envision a commercial system that replaces the driver box and laptop, both of which are currently carried in a backpack, with a single iPod-sized unit.

After examining the current state of transmitter/receiver technology, we have determined that only acoustic (and in particular ultrasonic) components offer high precision, low cost, and small size.

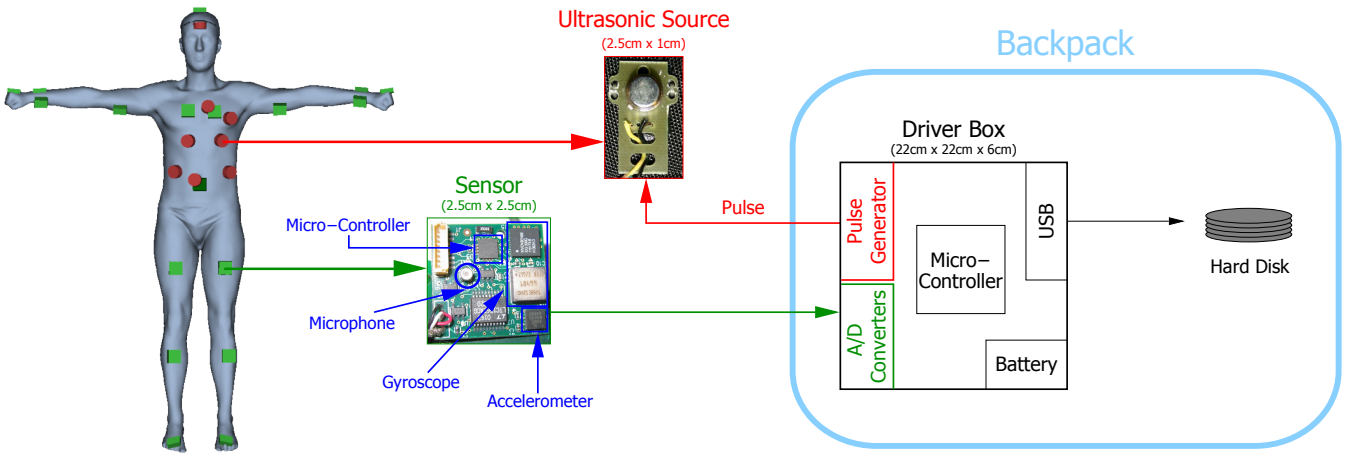

Figure 2: System Prototype. Our system consists of an array of small, low-cost, low-power ultrasonic sources and detectors (microphones) placed on the body (left). The ultrasonic sources (top-middle) sequentially emit ultrasonic pulses, which are received by the microphones (bottom-middle) and processed to yield distance measurements for all source-microphone pairs. To increase precision and the sampling rate, as well as to alleviate visibility problems, each sensor board is also equipped with a 3-axis gyroscope and a 3-axis accelerometer that measure rotation rates and linear accelerations respectively (bottom-middle). The data collection is managed by a small driver box (right) using a laptop hard disk for storage; both the driver box and the laptop are carried by the user in a backpack.

Therefore, our signal sources employ off-the-shelf piezoelectric transducers (Figure 2 top-center) to emit pulses at ultrasonic frequencies (40 kHz). They are mounted onto small plastic plates, attached to the garment, and wired to the pulse-generating driver box. The pulses are detected by conventional audio microphones (Figure 2, bottom-center). Although they do not exhibit optimal response in the 40 kHz range, they are able to clearly detect our ultrasonic pulses while offering several advantages over ultrasonic detectors: they are small in size (2.5mm^3); they have a wide-angle response — there is no need for accurate alignment with the sources; and they have a wide bandwidth — we do not need to tune them to the exact frequency of the ultrasonic source. We arranged the ultrasonic sources such that they see most of the microphones most of the time: seven sources around the chest and belly pointing forward, with the eighth source on the brim of a hat pointing down (Figure 2, left).

In addition to the microphone, each sensor board (Figure 2, bottom-center) is equipped with a 3-axis rotation rate sensing unit (the gyroscope), and a 3-axis linear acceleration sensing unit (the accelerometer). Their measurements enhance the precision and frame rate of the ultrasonic components. Furthermore, they alleviate the line-of-sight problems associated with acoustic signals. An on-board micro-controller collects the inertial data, combines it with the acoustic signal, and sends it to the driver box.

The driver box has three main tasks: to generate pulses that drive each ultrasonic source, sample the data from each of the sensor boards, and provide power to all inertial and ultrasonic components. As a result, all of our data is perfectly synchronized (we know exactly when the pulses are emitted with respect to each sensor signal). The sampling rate of the A/D converters in the driver box is about 150kHz, well above the Nyquist rate of the 40kHz ultrasonic pulses and the 13kbps inertial data (see below). In addition, the box houses a USB hub through which the sampled signals from each sensor board are transferred to a hard disk.

3.2 Ultrasonic Operation

Our ultrasonic subsystem operates similarly to a conventional acoustic ranging system, where there is a single source and a single detector. At regular intervals, the source emits a short burst of ultrasonic energy (a "pulse"), which is subsequently sensed by the detector. For example, our pulses are ten cycles wide at 40 kHz. The observed time delay ("time of flight") between the emission of the pulse and its detection is proportional to the distance between the two.

As the signal propagates through the air and bounces off objects in the environment, the detector will record several pulses at different times. The earliest detected pulse is the one that corresponds to the direct line-of-sight (LOS) and should be used to determine distance. The subsequent reflected pulses generally will be progressively weaker as they have to travel further through the air.

In our system, we also need to distinguish between pulses emitted by different sources. To accomplish this, the sources emit pulses at different times in a round-robin fashion (similarly to [Randell and Muller 2001]). The time separation between pulses from different sources must be long enough to ensure that reflected pulses from one source are not mistaken for the LOS pulse from the next source in the sequence. We have selected a conservative time interval of about 8 ms between the subsequent pulses. At the average speed of sound, this corresponds to a distance of about 2.75m the pulse will travel before another pulse is emitted by another source. We have found this to be sufficient to ensure that the LOS pulse is considerably stronger than any reflected pulse from a previous source. Since our system includes eight sources, each individual source emits pulses at 64 ms intervals.

The microphone on each of our sensor boards senses the ultrasonic pulses from all the visible ultrasonic sources. As the top row of Figure 3 visualizes, the corresponding analog signal is amplified and filtered in order to enhance its quality in the 40kHz range. The resulting analog signal, together with the digital inertial signal, is sent to the driver box and and stored on the laptop's hard disk.

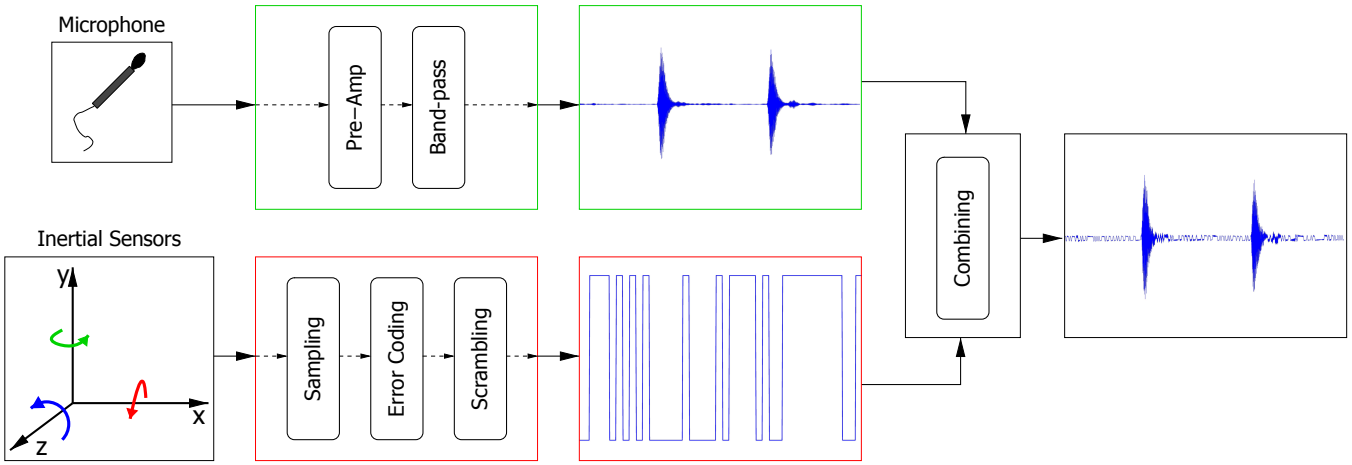

Figure 3: Sensor Operation. Our sensor boards combine acoustic and inertial data, and send the resulting signal to the driver box, which samples and stores it onto a hard disk. The acoustic signal sensed by the microphones (top) is amplified and filtered to enhance the quality of the ultrasonic pulses. At the same time, the inertial data from the gyroscopes and accelerometers (bottom) is digitally encoded by the on-board micro-processor, ensuring faithful reconstruction with error correction code and scrambling. The analog acoustic signal and the 13kHz digital inertial signal are multiplexed together and transmitted on a single wire.

3.3 Inertial Operation

Our inertial subsystem operates independently of the ultrasonic components. The gyroscopes and accelerometers measure rotational rates and accelerations. The micro-processor on each sensor board samples them as 10-bit quantities and accumulates several readings to obtain more precise 12-bit quantities. We increase the frame rate by not sending the internal 12-bit values. Instead, we turn them into 6-bit values using delta modulation to maintain good precision. In addition, we employ error-correction coding, enabling the amplitude of the digital data to be much lower than that of the acoustic signal and therefore causing less interference. Finally, we interleave and scramble our data with a pseudo-random sequence, which helps with error correction and prevents the baseline drift. The resulting values from the three gyroscope axes and the three accelerometer axes are encoded as a 13kbps digital stream and multiplexed with the analog acoustic signal. The sampling rate of the inertial data is 140Hz.

4 Signal Processing

In the signal processing stage, our system extracts distance and inertial measurements from the stored sensor signals. It extracts pulse locations from the sampled acoustic signals and converts them into distances. It also converts the digitally encoded inertial sensor voltages into accelerations and angular velocities. To obtain accurate measurements for both these steps, we perform precise calibration of our sensors. In the following section we overview both signal processing steps as well as the calibration procedure, and refer the reader to [Adelsberger 2007] for more details.

4.1 Ultrasonic Processing

We process the stored signal to extract distance measurements, noting that the microphone data are perfectly synchronized with the emitted pulses. Therefore, we can partition the signal into frames, where each frame corresponds to the maximum distance traveled by the pulse (2.75 m, which translates to 1120 samples). As visualized in the top row of Figure 4, we first band-pass the signal to eliminate all frequencies that are outside the range of our ultrasonic sources (these include the multiplexed 13kBps digital data). Based on the specifications of the ultrasonic source, we use a filter that is centered at 40kHz and has a width of 5kHz. Second, we square the signal, since we are more interested in its power than in the signal itself. Third, we extract the envelope of the signal power by applying a low-pass filter with a cut-off frequency of 30kHz. We observe that tracking the location of the peak does not provide the most precise distance measurement since the gradient of the signal power envelope is low. Instead, we compute the inflection point—the point where the gradient is the largest. In our case, this point is positioned about 40 samples after the start of the pulse. We perform a calibration for each ultrasonic source to compute the exact offset in the number of samples.

The power envelope of the signal can contain multiple peaks due to reflection. Furthermore, it can contain no useful peaks if there is no direct LOS between the source and the detector. Therefore, with each distance measurement we associate a confidence measure w ranging from 0 (no useful measurement) to 1 (a correct measurement). We represent the confidence measure w as a product of three factors: signal strength (w_s), temporal continuity (w_t), and an angle between the ultrasonic source normal and the detector (w_a).

To incorporate the signal strength factor, we first ensure that the signal is well above the estimated ambient noise, which we assume to be a zero-mean Gaussian. We also normalize the peak values by multiplying them by their squared distances since the signal strength of a spherical wave is proportional to the inverse of the squared radius. If the resulting value is above a predetermined threshold, then w_s is set to 1. Otherwise, w_s decreases with the squared inverse of the difference between the threshold and the peak value. The temporal continuity measure w_t is equal to 1 if the difference between the corresponding peak in two neighboring time instants is within a threshold, and decreases to zero as that difference grows. The angular confidence measure w_a is computed based on the current estimates of the sensor locations. In our implementation, w_a is set to 1 unless the angle between the ultrasonic source normal and the vector toward the sensor is greater than 90 degrees, in which case it is set to 0, ensuring that the microphone is within the field-of-view of the ultrasonic source. Figure 5 plots the confidence values for several source-sensor pairs throughout a

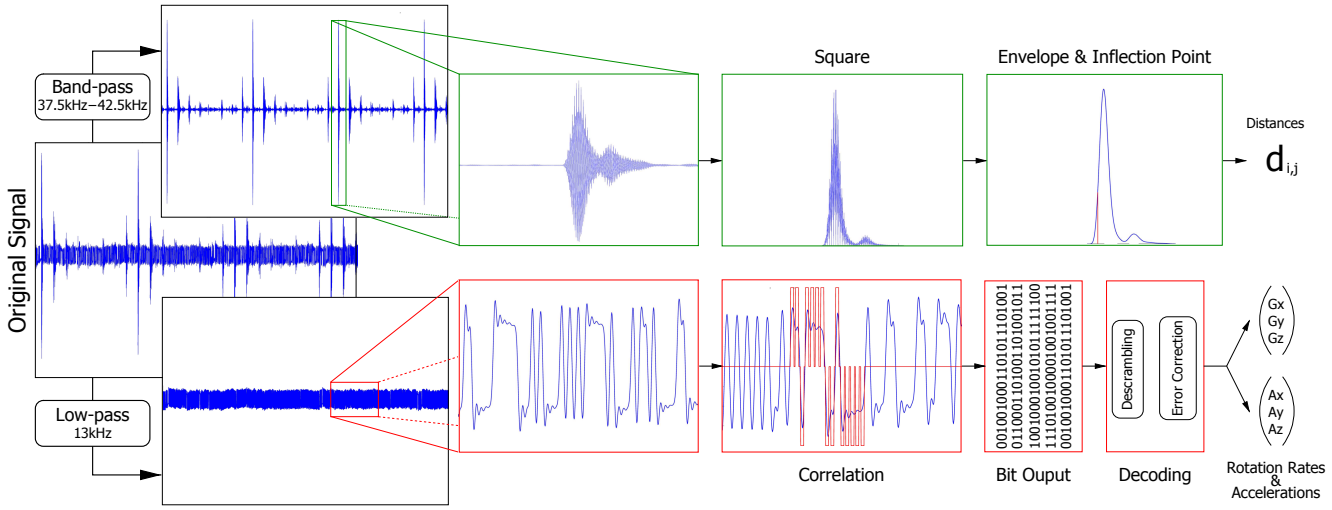

Figure 4: Signal Processing. The stored data is processed off-line to yield distance measurements from the analog acoustic data (top), and the rotation rates and accelerations from the digital inertial data (bottom). Top: After isolating the acoustic signal by band-passing the original signal around 40kHz, we compute its power envelope. The inflection points of the power envelope provide a robust indication of detected pulse locations, which in turn provide estimates of distances between ultrasonic sources and microphones. Bottom: The digital signal is isolated by low-passing below 13kHz, after which the scrambling pattern is used to identify the beginnings of different samples. The descrambling and error correction recover the encoded rotation rates and accelerations.

30-second motion exercising all joints. In general, the quality of the signal between each source-sensor pair varies with time, depending on the motion.

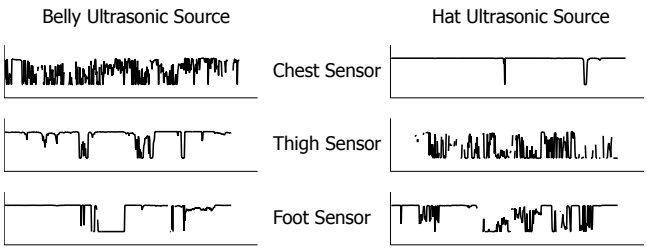

Figure 5: Confidence values for several source-sensor pairs are plotted over a 30-second motion exercising all joints. The belly source (left column) has a better line-of-sight with the thigh sensor (second row), while the hat source (right column) has a better view of the chest sensor (first row). They both see the foot sensor (third row) equally well but for different reasons: the belly source is positioned closer, while the hat source is better oriented.

4.2 Inertial Processing

Along with distances, we also extract the digitally encoded inertial data from the stored signal. As depicted in the bottom row of Figure 4, we first isolate the inertial portion of the signal by low-passing it with a cut-off frequency of 13kHz. We use the known scrambling pattern to lock onto the beginning of each new inertial sample and use thresholding to recover the bits of data. Those bits are then descrambled and error-corrected to yield the rotation rate readings for the three gyroscope axes, and the acceleration readings for the three accelerometer axes.

Even though the inertial readings are represented as 6-bit values, delta-modulation ensures that the overall precision is not degraded. For example, if the 6-bit round-off underestimates the acceleration in one frame, it will compensate by overestimating in the next frame; if we were to double-integrate the delta-modulated acceleration, the resulting position would have more than 6 bits of precision.

4.3 Calibration

The processed values from the ultrasonic signal correspond to the number of samples between the pulse emission and the detected inflection point, while the values from the inertial signal correspond to the voltages given by the inertial sensors. To convert these values to meaningful distances, accelerations, and rotation rates, we carefully calibrated all the components of our system.

For the ultrasonic components, we identified the offsets between the detected inflection points and the true pulse beginnings. These offsets differ from source to source, but are fairly constant across different microphones. This is because our microphones are of much better quality and sensing capabilities than our sources. We found the offsets by affixing the ultrasonic sources and microphones at several different locations (ranging from 30cm to 90cm apart), calculating the resulting inflection points, and measuring the actual distances using a FARO arm contact digitizer (*faro.com*). We obtained offsets for each ultrasonic source, yielding a distance error of 2.35mm mean and 1.92mm standard deviation according to our leave-one-out experiments.

For the accelerometers and gyroscopes, we identified zero crossings and slopes to convert their voltages to physical values assuming a linear model. Using a level, we aligned each axis of the accelerometer along the gravity and opposite gravity. Accumulating over a period of time, we obtained accurate estimates of $\pm g$, and the zero crossing as their mean. We then affixed each gyroscope to a turntable, orienting each of its axes both up and down. The turntable was rotating at 45rpm with 0.1% accuracy, enabling us to find the voltages corresponding to $\pm$45rpm. We averaged these two values to obtain the zero crossing.

5 Pose Estimation

Our system recovers body poses using angular velocities from gyroscopes, accelerations from accelerometers, and distances from the acoustic subsystem. While some approaches use algorithms specialized to only one kind of observation (e.g., [O'Brien et al. 2000; Theobalt et al. 2004; Kirk et al. 2005]), we employ the Extended Kalman Filter [Gelb 1974] to combine information from all three sensor types. Extended Kalman Filter (EKF) provides a convenient, efficient, and elegant framework for combining different types of measurements to recover the state of a given system [Welch and Bishop 1997]. It incorporates a model of the system dynamics with its indirect observations to yield pose estimates. On a high level, EKF evaluates the system dynamics to evolve the system state until the next observation, then uses this observation to improve its estimate of the system state.

5.1 System Dynamics

The body structure provides constraints that aid the recovery of its joint configuration, or pose. The pose of an articulated body is specified by the joint angles that describe the configuration of the shoulders, elbows, and other body joints. We use a single vector θ to assemble all joint angles in the body. This joint structure determines the forward-kinematics equations $\mathbf{F}(\theta)$ and $\mathbb{F}(\theta)$, which are used to compute position and orientation of any sensor.

The system state $\mathbf{x}$ contains joint angles, their velocities, and accelerations (θ, $\dot{\theta}$, $\ddot{\theta}$). Because we do not know the internal muscle forces, we assume that the change in accelerations between frame $k-1$ and frame k is governed by the zero-mean Gaussian noise $\mathbf{w}$:

$$\mathbf{x}_k = \begin{bmatrix} \theta \\ \dot{\theta} \\ \ddot{\theta} \end{bmatrix}_k = \begin{bmatrix} \theta + \dot{\theta} \\ \dot{\theta} + \ddot{\theta} \\ \ddot{\theta} + \mathbf{w} \end{bmatrix}_{k-1} = f(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}), \tag{1}$$

We hand tune the standard deviation of the noise term by tracking several motions. Setting this value to 0.04 rad/s^2 works well for most of our examples.

5.2 System Observations

Accelerometers provide acceleration readings in the local coordinate frame of each sensor. They sense the Earth's gravity as an upward acceleration of g even when the body is stationary. To derive sensor accelerations as a function of body joints, we first express the position of each sensor through forward kinematics as $p = \mathbf{F}(\theta)$. Differentiating with respect to time and applying the chain rule yields an expression for velocity $v = \mathbf{J}\dot{\theta}$, with $\mathbf{J} = d\mathbf{F}/d\theta$ being the positional forward-kinematics Jacobian. Differentiating once again, we calculate the accelerations as $a = \mathbf{J}\ddot{\theta} + \dot{\mathbf{J}}\dot{\theta}$. After rotating into the sensor coordinate frame, we express acceleration observations at frame k with the following function h:

$$\mathbf{z}_k = [z_i]_k = \left[\mathtt{rot}\left\{\dot{\mathbf{J}}_i\dot{\theta} + \mathbf{J}_i\ddot{\theta} - g\right\} + v_i\right]_k = h(\mathbf{x}_k, \mathbf{v}_k), \tag{2}$$

where $\mathtt{rot}\{\cdot\}$ denotes the rotation from the global coordinate frame to the coordinate frame of sensor i, and the standard deviation of the Gaussian noise $\mathbf{v}$ corresponds to the accelerometer precision of about 0.02m/s^2.

Gyroscopes measure angular velocity in the local frame of each sensor. To derive the angular velocity of each sensor as a function of body joints, we begin with the forward kinematics equation for the orientation quaternion: $q = \mathbb{F}(\theta)$. Taking a time derivative and applying the chain rule, we get $\dot{q} = \mathbb{J}\dot{\theta}$, where $\mathbb{J} = d\mathbb{F}/d\theta$ is the orientational forward-kinematics Jacobian. To get angular velocity, we multiply by the orientation quaternion conjugate and double the vector part of the resulting quaternion: $\omega = 2(q^*\dot{q})_{\mathtt{vec}} = 2(\mathbb{F}^*\mathbb{J}\dot{\theta})_{\mathtt{vec}}$. In the sensor coordinate frame, angular velocity observations at frame k are defined by the following function h:

$$\mathbf{z}_k = [z_i]_k = \left[\mathtt{rot}\left\{2(\mathbb{F}_i^*\mathbb{J}_i\dot{\theta})_{\mathtt{vec}}\right\} + v_i\right]_k = h(\mathbf{x}_k, \mathbf{v}_k) \tag{3}$$

where $\mathtt{rot}\{\cdot\}$ denotes the rotation from the global coordinate frame to the coordinate frame of sensor i, and the standard deviation of the Gaussian noise $\mathbf{v}$ corresponds to the gyroscope precision of about 0.002 rad/s.

The ultrasonic subsystem provides distances between a sensor i and a source j for all source-sensor pairs. The position of both the source and the sensor can be computed as a function of joint angles using the positional forward kinematics function $\mathbf{F}$. Since the distance observation timings are not synchronized with the inertial observations, we process at frame k all the distances that were measured between frame $k-1$ and k. The following function h expresses a set of distance observations in terms of the system state:

$$\mathbf{z}_k = \left[z_{ij}\right]_k = \left[\|\mathbf{F}(\theta)_i - \mathbf{F}(\theta)_j\| + v_{ij}\right]_k = h(\mathbf{x}_k, \mathbf{v}_k), \tag{4}$$

where the standard deviation of the Gaussian noise $\mathbf{v}$ corresponds to the ultrasonic subsystem precision of about 2.5mm, and is further divided by the confidence of each distance measurement.

5.3 Extended Kalman Filter

EKF incorporates the evolution of the underlying system state $\mathbf{x}$ along with the observations of the system. Each set of observations coming from accelerometers, gyroscopes, or the acoustic subsystem, is processed sequentially using the appropriate formulations of h and the corresponding measurement noise $\mathbf{v}$.

The Kalman time update step evolves the system until it reaches the next observation, yielding an *a priori* (before observation) estimate of the system state $\mathbf{x}$ and its covariance $\mathbf{P}$:

$$\mathbf{x}_k^- = f(\mathbf{x}_{k-1}, \mathbf{0}) \tag{5}$$

$$\mathbf{P}_k^- = \mathbf{A}_k\mathbf{P}_{k-1}\mathbf{A}_k^\top + \mathbf{W}_k\mathbf{Q}\mathbf{W}_k^\top, \tag{6}$$

where $^-$ stands for the *a priori* estimate, $\mathbf{Q}$ is the system noise covariance, $\mathbf{A} = \partial f/\partial \mathbf{x}$ is the Jacobian of f with respect to the system parameters, and $\mathbf{W} = \partial f/\partial \mathbf{w}$ is the Jacobian of f with respect to the system noise parameters.

The Kalman observation step uses the observation $\mathbf{z}_k$ to improve on the *a priori* estimates of the state $\mathbf{x}_k^-$ and its covariance $\mathbf{P}_k^-$:

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^\top(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^\top + \mathbf{V}_k\mathbf{R}_k\mathbf{V}_k^\top)^{-1} \tag{7}$$

$$\mathbf{x}_k = \mathbf{x}_k^- + \mathbf{K}_k(\mathbf{z}_k - h(\mathbf{x}_k^-, \mathbf{0})) \tag{8}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-, \tag{9}$$

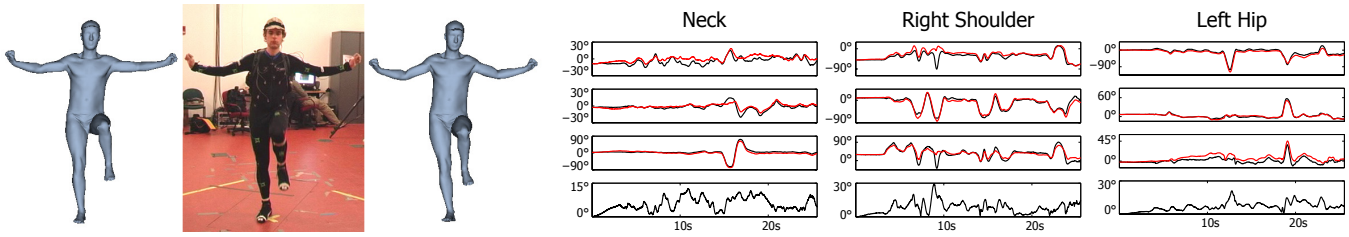

Figure 6: Comparison to the Vicon system. Left: one frame of the captured motion with our reconstruction to its left and Vicon reconstruction to its right. Right: a graph of a few 3-DOF joints (neck, right shoulder, and left hip) over 30 seconds of motion. The top three plots visualize the joint angles as reconstructed by us (red) and Vicon (black), while the bottom plot shows the orientation difference between our reconstruction and that of the Vicon system.

where $\mathbf{K}$ is the Kalman gain chosen to minimize the *a posteriori* state covariance, $\mathbf{R}$ is the measurement noise covariance, $\mathbf{H} = \partial h/\partial \mathbf{x}$ is the Jacobian of h with respect to the system parameters, and $\mathbf{V} = \partial h/\partial \mathbf{v}$ is the Jacobian of h with respect to the measurement noise parameters.

In our system, the noise covariance matrices ($\mathbf{Q}$, $\mathbf{R}$) are all diagonal, and their elements are the variances corresponding to the aforementioned standard deviations. All the Jacobian matrices ($\mathbf{A}, \mathbf{W}, \mathbf{H}, \mathbf{V}$) were computed analytically for speed.

5.4 Initialization

EKF is unstable if there are misalignments in the first frame because most of our measurements are incremental. We alleviate this problem by providing good estimates of the initial pose and the location of each body sensor. Our human mesh and its skeleton match the proportions of the subject, and the sensors are manually specified as rigid transforms in the coordinate frames of their parent bones. In addition, the subject begins each motion by holding a specified "rest" pose. As a result, the average accelerometer reading during that pose should be aligned with gravity (pointing upward). We exploit this fact to improve the pose of our model and tune the sensor orientations using a gradient descent approach. We additionally refine the inertial sensor offsets with an objective function which assures that the readings of each sensor during the initial pose integrate to zero.

6 Results

Our system is capable of acquiring motions ranging from biking and driving, to skiing, table tennis, and weight lifting (Figure 1). The accompanying video demonstrates its capability to reconstruct visible subtleties in recorded motions. The results are processed at a rate of 10 frames per second and visualized without any post-processing using an automatically skinned mesh [Baran and Popović 2007].

We evaluated the accuracy of our system by comparing it with Vicon's optical motion capture system, which is known for its exceptional precision. Since we could not collocate the optical markers with our sensors without interfering with the ultrasonic line-of-sight, we placed them around the body according to the suggestions in the Vicon manual. We used Vicon software to fit the same skeleton used by our system to the optical markers, and started all the reconstructions with an identical initial pose. To remove the effects of root drift in our reconstructions, we provided the root transform from Vicon's reconstruction to our Kalman filter at each frame. Although neither reconstruction is perfect, Vicon matches the original motions better, thus we treat it as ground truth in our analysis.

The left side of Figure 6 shows one frame of a 30-second motion, with our reconstruction on the left and Vicon's on the right. Qualitatively, Vicon's reconstruction matches the original motion better, although we are also able to reconstruct motion nuances such as slight wrist flicks. In addition, optical motion capture is able to recover the root transform without drift. The right side of Figure 6 shows plots of the neck, right shoulder, and left hip joints over the course of this motion. We visualize these three joints because they each have three degrees of freedom and are directly under the root in the skeletal hierarchy, and therefore free of parent-inherited errors. The top three plots for each joint compare the three Euler angles directly, with our reconstruction drawn in red and Vicon's in black. The bottom plot for each joint shows the orientation difference between the two reconstructions.

Our sensing capabilities have led us to explore multiple facets of our pose recovery system. We have reconstructed motions using various combinations of our sensors. Table 1 summarizes our findings for the same 30-second motion visualized in Figure 6. Compared to the Vicon reconstruction, the Extended Kalman Filter performs poorly with just accelerometers, just distances, or a combination of the two. Gyroscopes, on the other hand, provide much better reconstructions on their own, and yield even better results in conjunction with either the distances or the accelerations. The best results require using information from all available sensors. Although distance measurements do not have a dramatic effect on the reconstruction of activity in Table 1, we have observed significant drift in several experiments. For example, distance measurements are critical for an accurate reconstruction of the treadmill motion in Figure 7.

Limitations. Due to inherent physical limitations of our hardware components, we were unable to reconstruct high-impact motions such as jumping or hard kicking. Though sensors with wider sensing ranges are available to the detriment of precision, we have chosen ours to maximize the trade-off between coverage and precision. Other types of motions that we are unable to acquire with the current prototype include interaction between multiple subjects, such as dancing. By changing the ultrasonic source frequency of the partner, we could track each subject without interference, as well as obtain distance measurements between points on two different subjects.

Our distance measurements depend on the speed of sound, which is affected by temperature and, to a lesser extent, humidity. To obtain a more precise speed of sound, one could use a digital thermometer or a calibration device prior to each capture session. Distance measurements may also be affected by the presence of ultrasonic noise in the environment, but we have not experienced these problems in our experiments.

Sensors	Neck		Right Shoulder		Left Hip	
	$\mu(^\circ)$	$\sigma(^\circ)$	$\mu(^\circ)$	$\sigma(^\circ)$	$\mu(^\circ)$	$\sigma(^\circ)$
A	159.7	147.7	168.6	105.9	184.0	146.0
A+D	74.0	125.4	178.6	110.9	165.5	144.5
D	120.2	93.4	54.7	37.0	117.4	79.2
G	25.9	15.9	10.1	7.0	8.3	8.0
G+D	18.4	7.4	8.1	5.8	9.8	5.4
A+G	9.5	3.4	10.9	6.4	5.6	4.0
A+G+D	5.7	2.9	8.0	5.0	6.6	3.8

Table 1: Different combinations of our sensors (accelerometers **A**, gyroscopes **G**, and ultrasonic distances **D**) yield varying reconstruction quality as compared to the Vicon system. We report the mean μ and the standard deviation σ of the orientation differences between our reconstruction and Vicon's reconstruction for three 3-DOF joints of the skeleton.

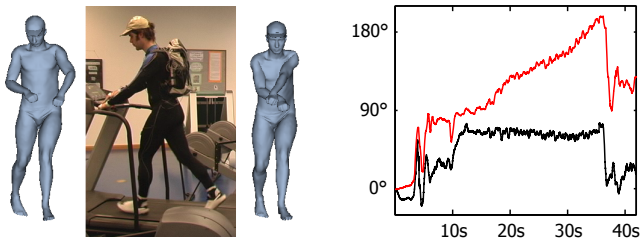

Figure 7: In the treadmill motion above, the distance measurements enabled us to avoid pose drift in the shoulder joint. Left: One frame of the treadmill motion where the reconstruction without distances (right) drifts, while the one with distances (left) does not. Right: A graph plotting the left shoulder joint angle for the reconstruction with (black) and without distances (red).

Perhaps the most significant limitation of our system is the lack of direct measurements of the root transformation. Our reconstructions exhibit drift in both global translation and rotation. We show that distance measurements help but they do not eliminate the problem entirely: over time, root drift propagates to other joints as well. Some of our experiments show that if we detect foot plants we can determine the global position and orientation relative to the ground plane. Another approach would be to use accelerometer readings for a vertical reference whenever the body is not accelerating. A more general solution could rely on additional sensors that perform absolute measurements (e.g., a GPS or magnetometer) to better constrain the root. These measurements can be incorporated into the EKF. Our evaluations, for example, incorporate root transforms obtained by the Vicon system. We could also compute absolute root measurements using ceiling-mounted ultrasonic sources [Foxlin et al. 1998] or image-based tracking. Lastly, optimization, which is slower than Kalman filtering, might converge to a better solution for the same sensor data.

7 Conclusions

We have presented a wearable motion capture system prototype that is entirely self-contained and capable of operating for extended periods of time in a large variety of environments. Our system acquires distance measurements between a set of ultrasonic sources and a set of sensors. Each sensor is augmented with inertial measurements in order to improve precision and sampling rate, as well as to alleviate line-of-sight problems. We have shown how to compute the pose of a human model directly from these measurements using the Extended Kalman Filter and qualitatively validated the performance of the system against a high-quality optical motion capture system. An attractive feature of our system is its low cost, with the current component price of around $3,000 (excluding the laptop). We believe that a much smaller version of the system with more sensors could be mass-produced for only a few hundred dollars, implying that this type of system could be owned and used on a daily basis by almost anyone.

There are many possibilities for future work. First, we could employ additional sensors to better constrain the root transform. Next, we could decrease the size and cost of our system. The driver box in particular has been built almost entirely from off-the-shelf components. By designing custom circuitry with only the necessary hardware, we could turn it into an iPod-like device that powers the system and stores the captured data, removing the need for a laptop. Another direction would be to perform all of the processing on-line, which might be achieved by accumulating several measurements before feeding them to the EKF to improve the speed. This would allow the system to be used as an input device in a variety of augmented-reality applications.

We should enrich motion repositories with varied data sets to further understand human motion. Restrictive recording requirements limit the scope of current motion data sets, which prevents the broader application of motion processing. An inexpensive and versatile motion-capture system would enable the collection of extremely large data sets. This enhanced infrastructure could then support large-scale analysis of human motion, including its style, efficiency, and adaptability. The analysis of daily human motion could even extend beyond computer graphics, and help prevent repetitive stress injuries, quicken rehabilitation, and enable design of improved computer interfaces.

8 Acknowledgments

We are grateful to Yeuhi Abe and Eugene Hsu for help with optical motion capture, Dragomir Anguelov for sharing the human mesh model, Ilya Baran for rigging the mesh, Edwin Olson for help with the EKF, Tom Buehler for video editing and proofreading, Emily Whiting for voice-over, and Jane Malcom for proofreading. This work was supported in part by the National Science Foundation (CCF-0541227) and software donations from Autodesk.

References

ADELSBERGER, R. 2007. *Practical Motion Capture in Natural Surroundings*. Master's thesis, ETH Zürich, Zürich, Switzerland.

BACHMANN, E. R. 2000. *Inertial and Magnetic Tracking of Limb Segment Orientation for Inserting Humans Into Synthetic Environments*. PhD thesis, Naval Postgraduate School, Monterey, California.

BARAN, I., AND POPOVIĆ, J. 2007. Automatic rigging and animation of 3d characters. *ACM Transactions on Graphics 26*, 3. In Press.

BISHOP, T. G. 1984. *Self-Tracker: A Smart Optical Sensor on Silicon*. PhD thesis, University of North Carolina at Chapel Hill.

BRASHEAR, H., STARNER, T., LUKOWICZ, P., AND JUNKER, H. 2003. Using multiple sensors for mobile sign language recognition. In *International Symposium on Wearable Computers*, 45–53.

BREGLER, C., AND MALIK, J. 1998. Tracking people with twists and exponential maps. In *Conference on Computer Vision and Pattern Recognition*, 8–15.

CHEN, Y., LEE, J., PARENT, R., AND MACHIRAJU, R. 2005. Markerless monocular motion capture using image features and physical constraints. In *Computer Graphics International*, 36–43.

DAVISON, A. J., DEUTSCHER, J., AND REID, I. D. 2001. Markerless motion capture of complex full-body movement for character animation. In *Computer Animation and Simulation*, 3–14.

FOXLIN, E., AND HARRINGTON, M. 2000. Weartrack: A self-referenced head and hand tracker for wearable computers and portable VR. In *International Symposium on Wearable Computers*, 155–162.

FOXLIN, E., HARRINGTON, M., AND PFEIFER, G. 1998. Constellation: A wide-range wireless motion-tracking system for augmented reality and virtual set applications. In *Proceedings of SIGGRAPH 98*, Computer Graphics Proceedings, Annual Conference Series, 371–378.

FOXLIN, E. 2005. Pedestrian tracking with shoe-mounted inertial sensors. *Computer Graphics and Applications 25*, 6, 38–46.

FREY, W. 1996. Off-the-shelf, real-time, human body motion capture for synthetic environments. Tech. Rep. NPSCS-96-003, Naval Postgraduate School, Monterey, California.

GELB, A., Ed. 1974. *Applied Optimal Estimation*. MIT Press, Cambridge, Massachusetts.

HAZAS, M., AND WARD, A. 2002. A novel broadband ultrasonic location system. In *International Conference on Ubiquitous Computing*, 264–280.

HIGHTOWER, J., AND BORRIELLO, G. 2001. Location systems for ubiquitous computing. *Computer 34*, 8 (Aug.), 57–66.

KIRK, A. G., OBRIEN, J. F., AND FORSYTH, D. A. 2005. Skeletal parameter estimation from optical motion capture data. In *Conference on Computer Vision and Pattern Recognition*, 782–788.

MEYER, K., APPLEWHITE, H. L., AND BIOCCA, F. A. 1992. A survey of position-trackers. *Presence 1*, 2, 173–200.

MILLER, N., JENKINS, O. C., KALLMANN, M., AND MATRIĆ, M. J. 2004. Motion capture from inertial sensing for untethered humanoid teleoperation. In *International Conference of Humanoid Robotics*, 547–565.

O'BRIEN, J., BODENHEIMER, R., BROSTOW, G., AND HODGINS, J. 2000. Automatic joint parameter estimation from magnetic motion capture data. In *Graphics Interface*, 53–60.

OLSON, E., LEONARD, J., AND TELLER, S. 2006. Robust range-only beacon localization. *Journal of Oceanic Engineering 31*, 4 (Oct.), 949–958.

PRIYANTHA, N., CHAKRABORTY, A., AND BALAKRISHNAN, H. 2000. The cricket location-support system. In *International Conference on Mobile Computing and Networking*, 32–43.

RANDELL, C., AND MULLER, H. L. 2001. Low cost indoor positioning system. In *International Conference on Ubiquitous Computing*, 42–48.

THEOBALT, C., DE AGUIAR, E., MAGNOR, M., THEISEL, H., AND SEIDEL, H.-P. 2004. Marker-free kinematic skeleton estimation from sequences of volume data. In *Symposium on Virtual Reality Software and Technology*, 57–64.

VALLIDIS, N. M. 2002. *WHISPER: a spread spectrum approach to occlusion in acoustic tracking*. PhD thesis, University of North Carolina at Chapel Hill.

WARD, A., JONES, A., AND HOPPER, A. 1997. A new location technique for the active office. *Personal Communications 4*, 5 (Oct.), 42–47.

WARD, J. A., LUKOWICZ, P., AND TRÖSTER, G. 2005. Gesture spotting using wrist worn microphone and 3-axis accelerometer. In *Joint Conference on Smart Objects and Ambient Intelligence*, 99–104.

WELCH, G., AND BISHOP, G. 1997. Scaat: Incremental tracking with incomplete information. In *Proceedings of SIGGRAPH 97*, Computer Graphics Proceedings, Annual Conference Series, 333–344.

WELCH, G., AND FOXLIN, E. 2002. Motion tracking: no silver bullet, but a respectable arsenal. *Computer Graphics and Applications 22*, 6 (Nov./Dec.), 24–38.

WOLTRING, H. J. 1974. New possibilities for human motion studies by real-time light spot position measurement. *Biotelemetry 1*, 3, 132–146.

YOKOKOHJI, Y., KITAOKA, Y., AND YOSHIKAWA, T. 2005. Motion capture from demonstrator's viewpoint and its application to robot teaching. *Journal of Robotic Systems 22*, 2, 87–97.