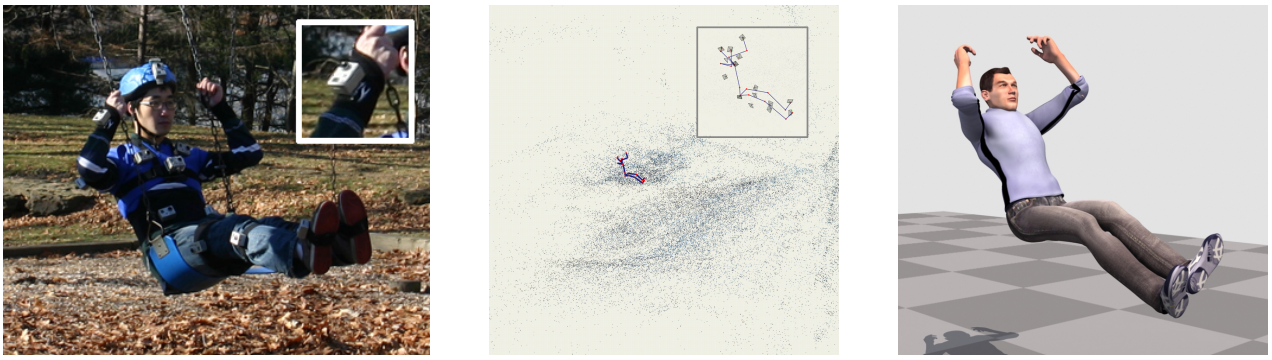
Motion Capture from Body-Mounted Cameras

Takaaki Shiratori* Hyun Soo Park† Leonid Sigal* Yaser Sheikh† Jessica K. Hodgins†*
* Disney Research, Pittsburgh † Carnegie Mellon University



(a) Body-mounted cameras (b) Skeletal motion and 3D structure (c) Rendered actor

Figure 1: *Capturing both relative and global motion in natural environments using cameras mounted on the body.*

Abstract

Motion capture technology generally requires that recordings be performed in a laboratory or closed stage setting with controlled lighting. This restriction precludes the capture of motions that require an outdoor setting or the traversal of large areas. In this paper, we present the theory and practice of using body-mounted cameras to reconstruct the motion of a subject. Outward-looking cameras are attached to the limbs of the subject, and the joint angles and root pose are estimated through non-linear optimization. The optimization objective function incorporates terms for image matching error and temporal continuity of motion. Structure-from-motion is used to estimate the skeleton structure and to provide initialization for the non-linear optimization procedure. Global motion is estimated and drift is controlled by matching the captured set of videos to reference imagery. We show results in settings where capture would be difficult or impossible with traditional motion capture systems, including walking outside and swinging on monkey bars. The quality of the motion reconstruction is evaluated by comparing our results against motion capture data produced by a commercially available optical system.

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

Keywords: Motion capture, structure-from-motion, articulated motion, wearable cameras

Links: DL PDF WEB VIDEO

*{shiratori, lsigal}@disneyresearch.com
†{hyunsoop, yaser, jkh}@cs.cmu.edu

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic motion for such scenes, the actors rode horses and robotic mockups in an expansive motion capture studio requiring a large number of cameras. Coverage and lighting problems often prevent directors from capturing motion in natural settings or in other large environments. Inertial systems, such as the one described by Vlasic and colleagues [2007], allow capture to occur in outdoor spaces but are designed to recover only the *relative* motion of the joints, not the global root motion.

In this paper, we present a wearable system of outward-looking cameras that allow the reconstruction of the relative and the global motion of an actor outside of a laboratory or closed stage. The cameras can be mounted on casual clothing (Figure 1(a)), are easily mounted and removed using Velcro attachments, and are lightweight enough to allow unimpeded movement. Structure-from-motion (SfM) is used to estimate the pose of the cameras throughout the capture. The estimated camera movements from a range-of-motion sequence are used to automatically build a skeleton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in traditional optical motion capture, because, in our system, any visually distinctive feature in the world can serve as a marker in the traditional systems. A by-product of the capture process is a sparse 3D structure of the scene. This structure is useful as a guide for defining the ground geometry and as a first sketch of the scene for 3D

animators and directors. We evaluate our approach against motion capture data generated by a Vicon optical motion capture system and report a mean joint position error of 1.76 cm and a mean joint angle error of 3.01° on the full range-of-motion sequence used for skeleton estimation. Our results demonstrate that the system can reconstruct actions that are difficult to capture with traditional motion capture systems, including outdoor activities in direct sunlight, activities that are occluded by near by proximal structures, and extended indoor activities.

Our prototype is the first, to our knowledge, to employ camera sensors for motion capture by measuring the environment and to estimate the motion of a set of cameras that are related by an underlying articulated structure. Current cameras are inexpensive, have form factors that rival inertial measurement units (IMUs), and are already embedded in everyday handheld devices. Our approach will continue to benefit from consumer trends that are driving cameras to become cheaper, smaller, faster, and more pervasive. Given the expected continuation of these technological trends, we believe that systems such as the one proposed here, will become viable alternatives to traditional motion capture technologies.

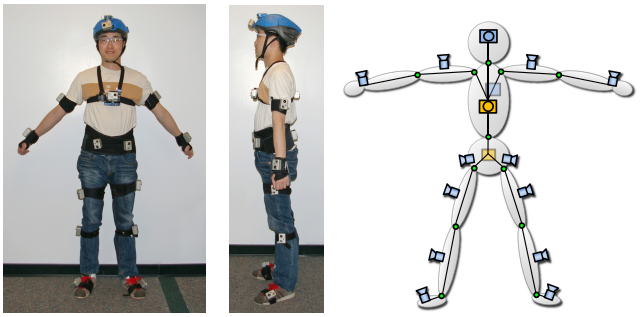
2 Related work

There are a variety of motion capture technologies currently available both commercially and as prototypes. The advantages and disadvantages of the different designs are discussed in several surveys (*e.g.*, [Welch and Foxlin 2002; Moeslund et al. 2006]). Motion capture systems can be classified as *outside-in* [Welch and Foxlin 2002], in that they rely on sensors mounted in the environment and passive, if any, markers on the body. By definition, this requirement restricts their use to laboratory environments or closed stage settings, because the capture space has to be instrumented with the sensors. *Inside-out* systems [Welch and Foxlin 2002] rely on sensors on the body to recover the 3D pose. This portability allows their use in both indoor and outdoor environments. Our approach falls into the latter category. Here, we review the most relevant methods and systems.

Optical motion capture systems [Woltring 1974] are among the most widely used in the industry today; commercial systems are available from Vicon (www.vicon.com) and Qualisys (www.qualisys.com), among others. Optical motion capture systems use a set of specialized high-resolution video cameras to track retro-reflective markers or light-emitting diodes (LEDs) placed at key points on the body. Triangulation is used to recover the 3D position of these markers in space, and the 3D marker positions, in turn, are used to fit a skeletal model to the observed motion. These systems are popular due to their accuracy; their major disadvantages are cost, portability, and intrusiveness. Optical systems require indoor setups that typically cost between tens and hundreds of thousands of dollars.

The use of photosensors was explored by Raskar and colleagues [2007]. Their proposed system relied on measuring the spatio-temporal light modulations produced by multiple LED transmitters that emitted gray coded patterns. The receiver modules, equipped with infrared and RGB photosensors, were tasked with decoding (demultiplexing) the observed patterns and, in doing so, directly producing the 3D spatial location (and as a side effect measuring incident light for scene light matching). While their system was inspirational for us in that it utilized a simplified photosensor as a "camera" worn on the body, it is fundamentally different from our approach, because it requires transmitters in the environment.

To alleviate the intrusive characteristics of marker-based motion capture systems, marker-less motion capture technologies have been developed by a number of researchers [Cheung et al. 2003; Deutscher and Reid 2005; Moeslund et al. 2006; Hasler et al. 2009]. Marker-less methods most often use regular video cameras with simple (*e.g.*, chromakey) backgrounds to reconstruct a voxel representation of the body over time and then fit a skeletal model to the voxel representations. A similar paradigm is used by the system developed by Organic Motion (www.organicmotion.com). Recent studies [Corazza et al. 2006; Corazza et al. 2010] suggest that with a sufficient number of cameras and favorable imaging conditions, the accuracy of marker-less methods can rival that of traditional optical motion capture. Hasler and colleagues [2009] introduced an approach to capture the motion of an actor in outdoor environments from multiple inward-looking moving cameras. The method uses audio to synchronize the cameras and fits a 3D scan of the actor to silhouettes estimated in each of the moving cameras. The marker-less methods require image segmentation, or a 3D scan of the actor.



(a) Front (b) Side (c) Skeleton

Figure 2: *Settings of cameras from (a) front view and (b) side view. (c) Illustration of skeleton and body-mounted cameras. Blue: cameras mounted on the body, and orange: cameras used as virtual cameras.*

The most direct approach to measuring human motion is through the use of a wearable electro-mechanical system; *e.g.*, Gypsy (www.animazoo.com). Such systems consist of an exoskeleton suit with embedded lightweight rods that articulate with the performer's bones. Potentiometers at the joints measure the angular rotation of the rods, and are converted to joint angles using a kinematic model. Such systems, while capable of directly measuring the motion of the subject, are intrusive and uncomfortable to wear.

Recently, there have been a number of self-contained, wearable experimental systems developed based on a variety of sensor technologies (*e.g.*, [Schwarz et al. 2010; Zhang et al. 2009]), including ultrasound, IMUs, and tri-axial accelerometers. Inertial motion capture systems (*e.g.*, Xsens MVN, www.xsens.com) measure the rotation of body parts in the world using accelerometers and gyroscopes. These systems are portable and can be taken outside; however, they are only able to measure the orientation of body parts, not the motion of the body in the world. Multiple sensors can be combined to alleviate drift. For example, Vlasic and colleagues [2007] added ultrasonic sensors to IMUs. Alternatives for battling drift include data-driven approaches based on motion capture data to stabilize accelerometer estimates [Slyper and Hodgins 2008; Xie et al. 2008; Kelly et al. 2010; Tautges et al. 2011].

Our system is camera-based and therefore relies on the rich data in a detailed view of the environment. We use the images from the cameras along with the estimated 3D geometry of the environment to recover the 3D limb positions and orientations in the world over time. Thus, we build on substantial prior work in SfM [Hartley and Zisserman 2004; Pollefeys et al. 2004; Snavely et al. 2006] and visual Simultaneous Localization and Mapping

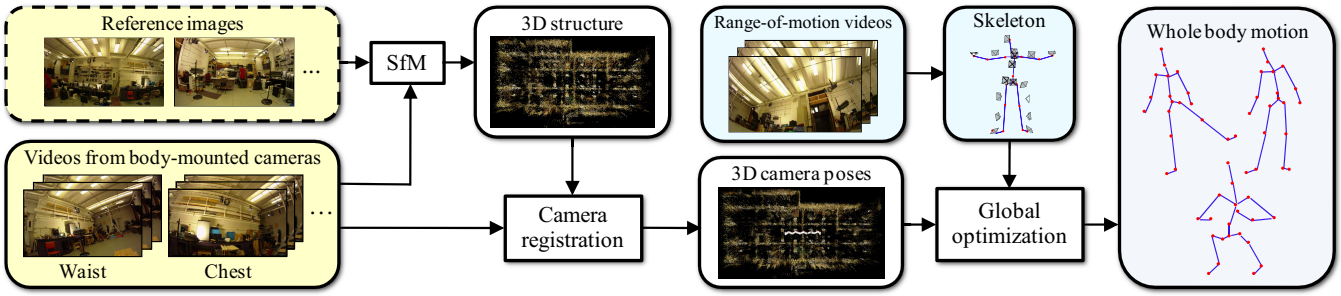


Figure 3: *Our system takes video data captured by body-mounted cameras and outputs the reconstruction of the human motion. The motion of the body is estimated by using SfM on individual cameras as an initial guess and optimizing the reprojection errors of the 3D structure while enforcing the underlying articulated relationships between cameras and the smoothness of motion across time.*

(SLAM) [Welch et al. 1999; Davison et al. 2007; Klein and Murray 2007]. These approaches have been used for estimating the motion of moving platforms [Ballan et al. 2010; Nístér et al. 2006] and even humans [Oskiper et al. 2007; Zhu et al. 2007; Zhu et al. 2008]. However, they recovered only the independent ego-motion of individual camera platforms. Our work is the first to reconstruct the 3D motion of a set of cameras related by an underlying articulated structure.

3 Hardware Setup

One camera is attached to each body segment using a Velcro strap-on mount, as shown in Figures 2(a) and (b). Three cameras are attached to the waist for root pose estimation, and two cameras are attached to the torso. The cameras are synchronized using a standard audio calibration signal[1]. The subject performs a range-of-motion trial for skeleton estimation and then performs the desired activity for capture. The video data are downloaded from the cameras after the capture for processing. Our system produces the skeleton of the actor, root position, and orientation and joint angles across time and also the 3D structure of the scene as a by-product.

We use 16 or more commercially available wide-angle (170° field of view) sport action cameras called HD Hero from GoPro (www.goprocamera.com) at a cost of 250 dollars per camera; making our entire setup approximately 5,000 dollars. The cameras are lightweight at 94 g and have a small form factor (42 mm × 60 mm × 30 mm). HD Hero cameras are equipped with a CMOS sensor and are capable of a variety of resolution/frame rate settings; we record at 720p (1280 × 720) resolution at 60 frames per second. If cameras on some body segments are often occluded by limbs (*e.g.*, waist, torso), we use additional cameras to provide robustness by creating a wider aggregate field of view.

All the cameras are calibrated in advance using a fisheye lens distortion model [Devernay and Faugeras 2000] to provide estimates of focal length, principal point, and the distortion coefficient. As the lens and focal length are fixed for the cameras, these estimates need to be computed only once and are re-usable across captures.

4 Reconstructing Human Motion

Conventional SfM can provide visually feasible estimates of 3D structure and camera pose, but these estimates are often not sufficiently accurate for capturing human motion. In order to compute appropriate estimates of human motion across time, our SfM solution considers the articulation of body-mounted cameras with the underlying skeleton of the actor and fits them to image measurements:

$$\{\mathcal{O}^*, \mathcal{A}^*\} = \underset{\mathcal{O},\mathcal{A}}{\operatorname{argmin}} \; E_r + \lambda_{\mathcal{O}} E_{\mathcal{O}} + \lambda_{\mathcal{A}} E_{\mathcal{A}}, \qquad (1)$$

where $\mathcal{O}$ and $\mathcal{A}$ are the time-series data of the root position and the joint angles, respectively. E_r accounts for reprojection errors of the 3D reconstruction with measured image feature locations using the skeleton constraint for the cameras. E_O and E_A consider the smoothness of resulting motion. $\lambda_{\mathcal{O}}$ and $\lambda_{\mathcal{A}}$ are weights that control the influence of the smoothness constraints.

Equation (1) is highly non-linear and optimization of the equation requires good initial estimates of camera poses and skeleton. We develop the pipeline shown in Figure 3. After the data are captured, we reconstruct the 3D structure of the scene from reference images using SfM. While this step is optional in principle, it substantially reduces the drift in the reconstructed motions, and we chose to perform it for all our captures. The 3D structure is used to reconstruct body-mounted camera poses across time (Section 4.1). If a new skeleton is required, the subject is asked to perform a standard range-of-motion exercise at the beginning of the capture session. The skeleton is automatically generated (see Appendix) and is used to reconstruct whole body poses from the cameras (Section 4.2). The user can optionally refine the skeleton by changing the pose of the camera with respect to the joint through a graphical user interface. Finally, the motion is refined using an image-based non-linear optimization that incorporates temporal smoothing (Section 4.3).

4.1 Initializing Camera Poses Using SfM

Direct incremental SfM from body-mounted cameras yields precise 3D reconstruction locally but suffers from global drift when the capture area is large and 3D structure is far from the camera locations [Hartley and Zisserman 2004]. To avoid this problem, we record *reference images* of the capture area, and reconstruct the 3D structure using the images. Using this 3D structure and corresponding 2D measurements from a body-mounted camera, the camera pose can be reconstructed. We call this process *absolute camera registration*. If there are significant differences in view between the reference images and the recorded videos, some cameras may not be reconstructed. We handle this situation by adding new structure points with newly registered cameras, and rerun the camera registration. We call this iterative process *relative camera registration*, and repeat the process until most of cameras are reconstructed.

3D Reconstruction of Reference Images: From the reference images, we extract Scale-Invariant Feature Transform (SIFT) key-

[1] We use a clapper board to produce a loud clap at the beginning and end of each trial. We find peaks in the audio signal of resulting movie files from all the cameras and look for the most consistent duration between peaks using a simple form of clustering and exhaustive search [Hasler et al. 2009].

(a) 3D structure (b) After absolute camera registration (c) After fifth iteration of relative camera registration

Figure 4: *(a) Reference structure reconstruction, (b) after absolute camera registration, and (c) after the fifth iteration of relative camera registration. Adding points from absolute and relative camera registration processes allows us to make the camera registration denser.*

points [Lowe 2004] and find correspondences between a pair of images using an approximate nearest neighbor search [Muja and Lowe 2009]. The fundamental matrix estimation based on RANSAC [Fischler and Bolles 1981] enables us to obtain geometrically consistent matches.

To estimate the extrinsic parameters of the cameras, we choose an initial pair of images that has a significant number of matches that cannot be accounted for by a homography. From those matches, we estimate the relative camera orientation and translation extracted by the essential matrix and triangulate the location of the matched feature points in 3D using the Direct Linear Transform algorithm [Hartley and Zisserman 2004], followed by a two-image bundle adjustment [Lourakis and Argyros 2009]. We incrementally add an image that has the greatest number of inlier 3D-2D correspondences, among the remaining images. From these correspondences, we reconstruct the camera pose using a Perspective-n-Point (PnP) algorithm [Lepetit et al. 2009] inside a RANSAC procedure.

Once the extrinsic parameters for the new camera are reconstructed, 2D-2D correspondences between reconstructed images and the newly added image are reconstructed in 3D. For accuracy, we exclude 3D points with the following criteria: any point that has high reprojection error ($>$1 pixel) and any point when the angle subtended by the rays used for triangulation is small ($<2^\circ$). Once the structure has been updated, a sparse bundle adjustment is run to refine the entire model. This process continues until most of the reference images are registered.

Absolute Camera Registration: After the 3D structure is reconstructed from the reference images, it is used to estimate the body-mounted camera poses. The process of registering images from body-mounted cameras is similar to that of adding a new reference image in the SfM process. Using RANSAC with PnP, we find the best extrinsic camera parameters that produce less than 1 pixel reprojection error when the number of inlier 3D-2D correspondences is sufficient ($>$50). Once the camera parameters are estimated, new 3D points are triangulated using 2D-2D correspondences between the newly registered image and the previously registered images. To reduce the computational cost of keypoint matching for new 3D points, we ignore camera pairs whose optical axes have more than 90° orientation difference. The criteria for adding a new point are the same as those used in the SfM process. The bundle adjustment refines newly registered camera poses and the 3D structure. The 3D structure obtained from the reference images is fixed during the optimization so that the structure can act as an anchor to avoid drift.

Relative Camera Registration: The reconstruction from the absolute camera registration may be sparse, particularly when the viewing angles of the reference images are different from those of images from the body-mounted cameras. To increase the density of the reconstruction for the body-mounted camera poses, we find matches between the images from the absolute-registered camera and the images from the unregistered cameras. Because the unregistered cameras are close to the absolute-registered cameras, the viewpoints are similar. This process enables us to reconstruct the poses of the remaining cameras.

The relative camera registration processes are iterated until camera registration is satisfactory. Figures 4(b) and 4(c) show the results of absolute and relative camera registration. While the absolute camera registration produces gaps, the fifth iteration of the relative camera registration fills most of the gaps.

Homographies for Unregistered Cameras: After the iterative camera registration, there may still be unregistered cameras for particular windows of time. This situation occurs, for example, when an actor performs a fast motion such as running and the images are blurry.

To deal with the remaining unregistered cameras, we estimate relative camera orientation between consecutive frames $\mathcal{C}_1$ and $\mathcal{C}_2$ using a homography. When the camera centers of two images coincide, the relative orientation can be estimated from the homographies. Here, we assume that the camera center difference between two consecutive frames is small enough to neglect, compared to the distance between the 3D points and the camera centers. We extract 2D-2D matches based on the SIFT keypoint descriptors and robustly find the consistent homography using RANSAC. Once the homography $\mathbf{H}$ is estimated, the relative orientation, ${}^{\mathcal{C}_2}\mathbf{R}_{\mathcal{C}_1}$, can be obtained by

$$ {}^{\mathcal{C}_2}\mathbf{R}_{\mathcal{C}_1} = \mathbf{K}_{\mathcal{C}_2}^{-1}\mathbf{H}\mathbf{K}_{\mathcal{C}_1}, \tag{2} $$

where $\mathbf{K}$ is an intrinsic parameter matrix. To avoid drift caused by one-way camera orientation estimation with homographies, we take an average of forward and backward interpolation. If camera positions are also needed, linear interpolation of the positions between registered cameras is used. This interpolation provides the initialization of joint angles and root positions. The inlier 2D-2D correspondences used for the homography computation are used as image measurements in the subsequent optimization.

4.2 Mapping Cameras to a Skeleton

At the beginning of the capture, the actor is asked to perform a predefined range-of-motion exercise, in which he exercises each joint through its full range of motion. We extract the underlying skeleton structure from the images recorded during the range-of-motion performance. As is common with commercial motion capture systems like Vicon, we use a predefined kinematic structure. One or more cameras are associated with each link in the kinematic structure, as shown in Figure 2(c). The root of the skeleton has six degrees of freedom (DOFs), and the joints have three DOFs. We apply the method of O'Brien and colleagues [2000] to estimate the skeleton and the 3D spatial relationship of each camera to the kinematic structure (see Appendix). We do not currently consider biomechanical constraints.

Forward Kinematics from Camera Poses: The skeleton provided by the range-of-motion exercise is parameterized by the root position, root orientation and joint angles. The root position and orientation are taken to be coincident with the root camera. Hence given the skeleton, we can obtain a pose for each time instant by applying the waist camera pose to the root segment directly and applying the relative orientations between pairs of cameras, along the kinematic chain, to the joints. Note that positions of the camera poses are not used except for the waist cameras.

Equation (1) considers the skeleton as a hard constraint for refinement. Forward kinematics enables us to maintain this constraint by estimating camera positions with respect to the skeleton. The Euclidean transformation from the joint coordinate system, $\mathcal{J}$, to the world coordinate system, $\mathcal{W}$, is defined as

$$ {}^{\mathcal{W}}\mathbf{T}_{\mathcal{J}}(t) = \begin{bmatrix} {}^{\mathcal{W}}\mathbf{R}_{\mathcal{J}}(t) & {}^{\mathcal{W}}\widetilde{\mathbf{p}}_{j}(t) \\ \mathbf{0} & 1 \end{bmatrix}, \tag{3} $$

where ${}^{\mathcal{W}}\mathbf{R}_{\mathcal{J}}$ and ${}^{\mathcal{W}}\widetilde{\mathbf{p}}_{j}$ are the orientation of the corresponding camera and the position of the joint in $\mathcal{W}$, respectively[2]. Therefore, the position of its child joint, ${}^{\mathcal{W}}\mathbf{p}_{j+1}(t)$, in $\mathcal{W}$ is computed as

$$ {}^{\mathcal{W}}\mathbf{p}_{j+1}(t) = {}^{\mathcal{W}}\mathbf{T}_{\mathcal{J}}(t)\, {}^{\mathcal{J}}\mathbf{q}, \tag{4} $$

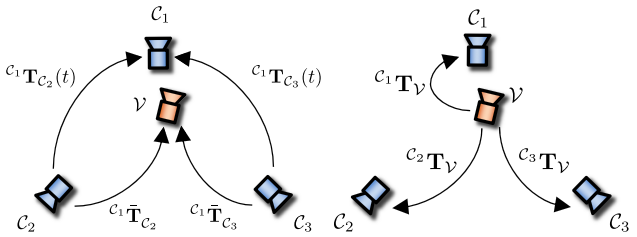
where ${}^{\mathcal{J}}\mathbf{q}$ is a vector from the parent joint to the child joint in $\mathcal{J}$. This formulation allows us to estimate the hierarchical joint position, recursively. Similarly, the camera position in the world coordinate system can be re-estimated as

$$ {}^{\mathcal{W}}\mathbf{C}_{j}(t) = {}^{\mathcal{W}}\mathbf{T}_{\mathcal{J}}(t) \begin{bmatrix} -{}^{\mathcal{J}}\widetilde{\mathbf{p}}_{j} \\ 1 \end{bmatrix}, \tag{5} $$

where ${}^{\mathcal{W}}\mathbf{C}_{j}(t)$ is a camera center attached to the j-th joint at time t in $\mathcal{W}$.

Virtual Cameras for Robust Limb Pose Estimation: Estimating body-attached camera poses from SfM while the body is moving is sometimes difficult because of motion blur, the rolling shutter effect, occlusion by limbs, and lack of texture in the background (*e.g.*, sky). Under such conditions, the camera poses cannot be reconstructed or are very noisy. When camera poses are mis-estimated, the resulting motion of the skeleton is incorrect. To alleviate this problem, we attach multiple cameras to the limb and estimate the limb motion from a *virtual camera*, which takes a robust average of those cameras (estimated using SfM). We use a virtual camera where occlusion occurs frequently, where a precise estimation is essential (*e.g.*, for the root), or where camera registration is difficult (*e.g.*, for the chest to account for non-rigidity, or shin to deal with fast motion and impacts that result in imaging artifacts). The virtual camera reduces the occlusion problem significantly and allows skeletal motion to be reconstructed robustly.

[2] $\widetilde{\mathbf{p}}$ is an inhomogeneous representation of $\mathbf{p}$.



(a) Computing a virtual camera pose with average transform

(b) Parameterizing physical cameras with a virtual camera

Figure 5: *Illustration of a virtual camera (blue) created from physical cameras (orange). The virtual cameras are used to combine information from multiple body-attached cameras for additional accuracy. (a) Average relative transforms between the cameras are estimated across time, and the virtual camera pose is estimated by applying the average relative transforms to the physical cameras at each time instant. (b) The physical cameras are parameterized by the virtual camera using the average relative transforms.*

The virtual camera poses can be estimated from motion over time. Here, we assume that there are three physical cameras, $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$, tightly connected to a single limb. One camera, *e.g.*, $\mathcal{C}_1$, is selected as a reference camera. As shown in Figure 5(a), the average relative transforms from the other two cameras to the reference camera[3], ${}^{\mathcal{C}_1}\bar{\mathbf{T}}_{\mathcal{C}_2}$ and ${}^{\mathcal{C}_1}\bar{\mathbf{T}}_{\mathcal{C}_3}$, can be estimated across time,

$$ {}^{\mathcal{C}_1}\bar{\mathbf{T}}_{\mathcal{C}_2} = f_a\left({}^{\mathcal{C}_1}\mathbf{T}_{\mathcal{C}_2}(1), {}^{\mathcal{C}_1}\mathbf{T}_{\mathcal{C}_2}(2), \cdots, {}^{\mathcal{C}_1}\mathbf{T}_{\mathcal{C}_2}(T)\right), \tag{6} $$

where $f_a(\cdot)$ is a function that takes an average of the transforms. Once the average transform is estimated, the inverse of the average transform is used as a transform from the virtual camera, $\mathcal{V}$, to the physical cameras, *i.e.*,

$$ {}^{\mathcal{C}_2}\mathbf{T}_{\mathcal{V}} = {}^{\mathcal{C}_1}\bar{\mathbf{T}}_{\mathcal{C}_2}^{-1}, \;\; {}^{\mathcal{C}_3}\mathbf{T}_{\mathcal{V}} = {}^{\mathcal{C}_1}\bar{\mathbf{T}}_{\mathcal{C}_3}^{-1}. \tag{7} $$

Then, the virtual camera pose can be obtained by again taking an average of the transforms for $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$.

Now the transforms from the virtual camera to each physical camera are known, which implies all physical cameras can be parameterized by the virtual camera pose (Figure 5(b)). This parameterization will be used when reprojection errors are computed in the subsequent optimization.

4.3 Estimating Body Poses with Global Optimization

The final step is to optimize body poses $\mathcal{O}$ and $\mathcal{A}$ by minimizing the objective function in Equation (1). Conceptually, the optimization seeks to find body poses of the skeleton, over time, that are temporally smooth and result in low spatial error between the projected 3D structure, through the estimated cameras, and the actual observed structure in the images. The initial guess of the body pose is set with the registered camera poses and homographies, and the Levenberg-Marquardt method is applied to refine the poses. Considering all poses over time in the optimization is computationally expensive. Instead, we use a short time window and sequentially optimize the poses by shifting the window.

[3] If the cameras are rigidly connected, the average relative transforms are exactly the same as the relative transform at each time instant.

Reprojection Error Term: E_r refines whole body poses based on the reconstructed 3D structure and image measurements:

$$E_r = \sum_{j,t,p} \|P_j\left(\mathbf{X}_p, t, \mathcal{O}, \mathcal{A}\right) - \mathbf{x}_{j,t,p}\|^2_\Sigma + \sum_{j,t,h} \|H_j(t, \mathcal{A}, \widehat{\mathbf{x}}_{j,t,h}) - \widehat{\mathbf{x}}_{j,t-1,h}\|^2_\Sigma, \quad (8)$$

where $P(\cdot)$ is a camera projection function, $H(\cdot)$ is a function to apply a homography between consecutive images to an image measurement, and j, t, p, and h are indices of cameras, time, 3D points and 2D measurements for homographies, respectively. $\mathbf{X}$ is the location of the 3D structure point in the world coordinate system, and $\mathbf{x}$ is the corresponding 2D measurement. $\widehat{\mathbf{x}}$ is a 2D measurement after lens distortion correction for the homography.

The first term considers the reprojection errors of the 3D points with the 2D measurements for the registered cameras. This minimization is different from typical bundle adjustment of SfM in that the camera poses are constrained by the skeleton. Using the projection matrix, $\mathbf{P}_j(t)$, of the camera associated with the j-th joint, the projection function P_j is represented as

$$P_j\left(\mathbf{X}_p, t, \mathcal{O}, \mathcal{A}\right) = L_j\left(\frac{\mathbf{P}_{j:1}(t)\mathbf{X}_p}{\mathbf{P}_{j:3}(t)\mathbf{X}_p}, \frac{\mathbf{P}_{j:2}(t)\mathbf{X}_p}{\mathbf{P}_{j:3}(t)\mathbf{X}_p}\right), \quad (9)$$

where $L_j(\cdot)$ distorts the reprojected position using the fisheye lens distortion parameter of the j-th camera, and $\mathbf{P}_{j:i}$ is the i-th row of the projection matrix $\mathbf{P}_j$.

The second term is for the cameras that cannot be registered through the absolute and relative registration. The rotation matrices of the homographies in Equation (2) are parameterized with the joint angles. The inlier 2D-2D correspondences detected in the RANSAC-based homography estimation are used as image measurements.

Smoothness Terms: $E_\mathcal{O}$ and $E_\mathcal{A}$ can be also considered to obtain smooth motion. The differences of the root positions and joint angles between consecutive frames are minimized as

$$E_\mathcal{O} = \sum_t \|\mathcal{O}(t) - \mathcal{O}(t-1)\|^2_\Sigma, \quad (10)$$

$$E_\mathcal{A} = \sum_t \|\mathcal{A}(t) - \mathcal{A}(t-1)\|^2_\Sigma. \quad (11)$$

These terms are effective, particularly when the camera poses estimated from the absolute and relative registration contain undesirable jitter.

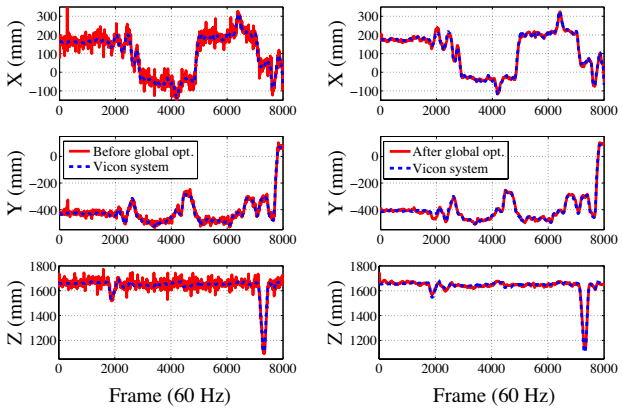
5 Results

In this section, we evaluate our system quantitatively using a conventional motion capture system as ground truth, and show additional results collected out of doors.

5.1 Quantitative Evaluation

First, we evaluate the effect of the global optimization step. Figure 6(a) shows the comparison between the camera centers estimated by the Vicon markers and our reconstruction before the global optimization but after the camera centers were adjusted based on the estimated skeleton. Figure 6(b) shows the reduction in error after global optimization with the smoothness terms.

Figure 7(a) compares the joint angle trajectories obtained by our system with the measurements from the Vicon motion capture system. The top row shows the joint angle trajectories of the upper body and the bottom row shows the joint angle trajectories of the lower body. The joint angles illustrated in the figure are the angle of the axis-angle representation normalized by the angle of the first frame in the capture session. The mean and median errors are 3.0093° and 1.8076°, respectively, and the minimum and the maximum errors are 0.038° and 9.52°, respectively. The standard deviation is 2.1891°. Because the error of a parent joint angle propagates to a child joint, the joint angle errors may not be sufficient to characterize the error of the overall system. Therefore, we also evaluate the errors of the joint positions (Figure 7(b)). The error does not propagate significantly, because the optimization of Equation (1) finds a solution such that all cameras satisfy the image measurements. The mean and median position errors are 1.76 cm and 1.42 cm, respectively, and the minimum and the maximum errors are 0.053 cm and 12.24 cm, respectively. The standard deviation is 1.26 cm.



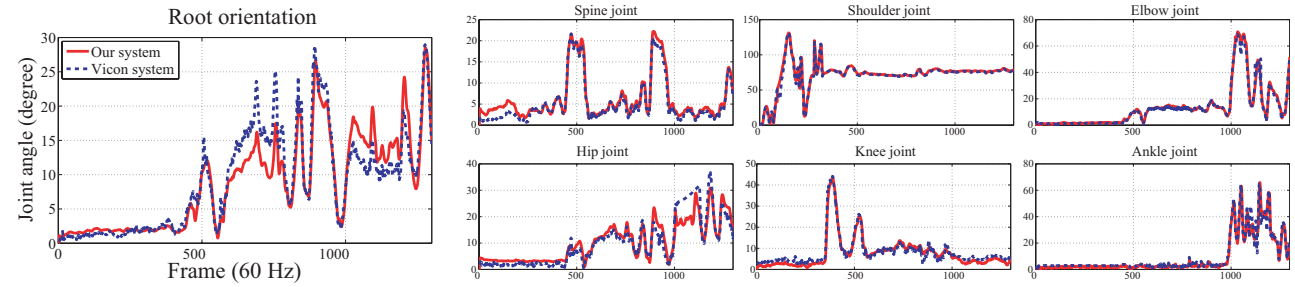
(a) Before global optimization (b) After global optimization

Figure 6: *Quantitative comparison of estimated camera centers with those obtained using a motion capture system. (a) SfM produces a noisy reconstruction of the camera poses, and (b) the non-linear optimization with the smoothness terms results in a more accurate estimation.*

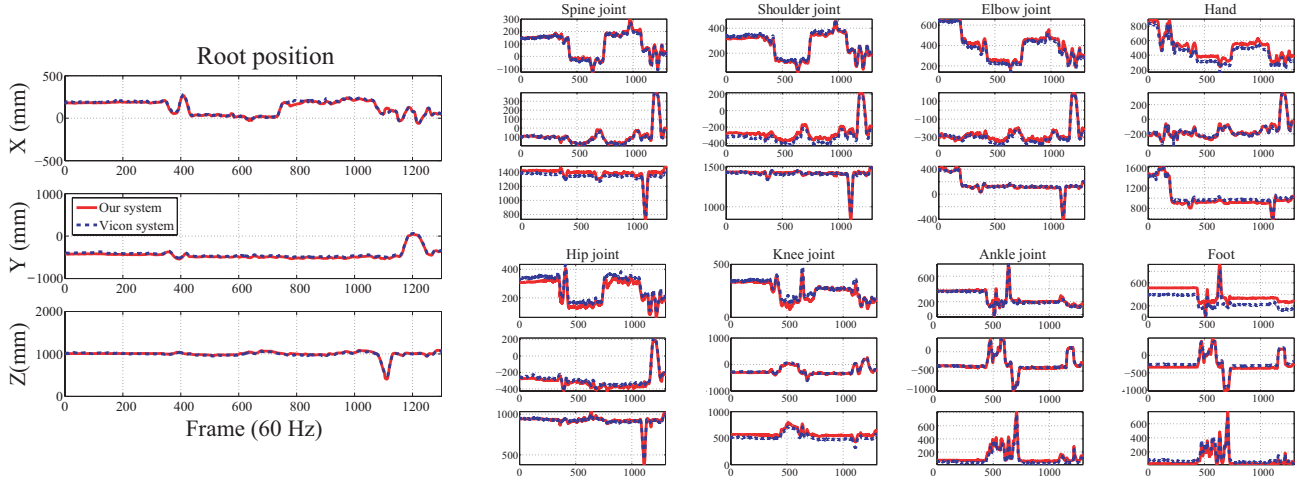
Method of Comparison: We now describe how we obtained these quantitative comparisons. Our system produces camera poses in the SfM space, while the motion capture system outputs 3D marker positions in the motion capture space. To compare the two different reconstructions, we needed to compute the following transforms between the two spaces.

We attached three markers on each of the cameras and several markers on static objects and collected images from the cameras and the corresponding marker positions from the motion capture system as the subject moved. Using the 3D positions of the static markers in the motion capture space and the corresponding image measurements specified manually, the camera center positions and orientations in the motion capture space were estimated. Thus we could convert the three marker positions in the motion capture data to the camera poses.

To recover the similarity transform from the SfM space to the motion capture space, we estimated a scale from the distances between the camera center pairs in both of the spaces. Then, we estimated translation and orientation from the SfM space to the motion capture space by applying the iterative closest point algorithm to the two sets of the camera centers. The parameters were used for the similarity transform after non-linear refinement.



(a) Joint angle comparison



(b) Joint position comparison

Figure 7: *Comparison of (a) joint angle trajectories and (b) joint position trajectories with the motion measured by the Vicon motion capture system.*

5.2 Outdoor experiments

The major benefit of our system is that it is portable and self-contained, allowing prolonged captures in outdoor environments. To illustrate these benefits we captured two sequences in the local playground; the results are illustrated in Figures 8(a) and 8(b). We also tested the ability to capture fast motions with a running motion on a street (Figure 8(c)). The top rows show the photos of the subject performing the motions, and the bottom rows illustrate a posed skinned character using the joint angles estimated by our system. Some of these motions were quite dynamic and we observed faster than 2.4 m/s instantaneous velocity of a camera in the swing and running sequences. Though these motions resulted in image blur, and the rolling shutter effect, we were able to properly reconstruct the sequences.

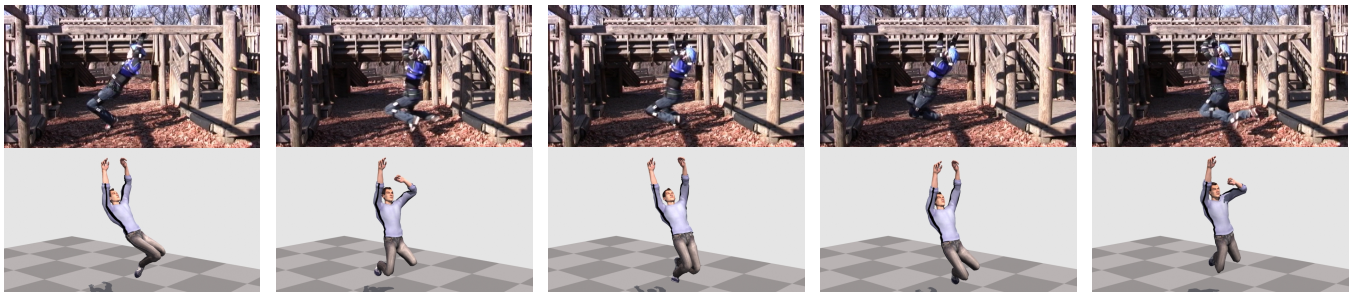
Figure 9 shows a reconstructed long walking motion along the winding path on an uneven terrain. The subject traversed a considerable distance that is far greater than what would be possible in a traditional indoor motion capture setup. We superimposed the sparse 3D structure and manually matched the viewing angle to a photo taken during the capture for reference. The sparse structure provides the context for the motion by showing the path along which the subject has walked.

6 Discussion

We introduce a novel system for capturing human motion in both indoor and outdoor environments. Our system consists of 16 or more consumer video cameras attached to body segments. We estimate their motion with respect to the world geometry through a SfM algorithm. We then relate and refine camera and skeletal motion through a non-linear optimization procedure. Our system has a number of advantages over traditional optical and IMU-based systems, because it: (i) requires no instrumentation of the environment and can easily be taken outside, (ii) is stable and does not suffer from drift, and (iii) provides sparse 3D reconstruction of the world for contextual replay or scene creation.

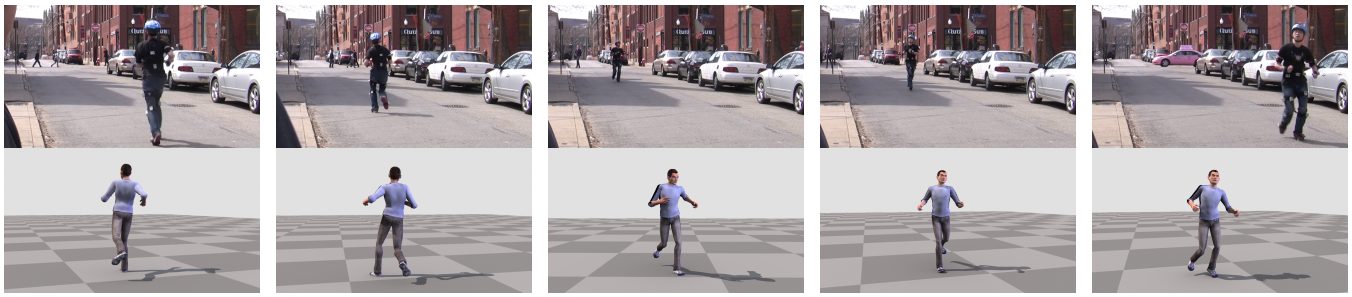
The principal causes of failure for our system are motion blur, automatic white balancing, rolling shutter effects, and motion in the scene. Low light and the cropped-frame formats found in many commercially available cameras can introduce motion blur as the camera moves quickly. The blur makes it difficult to estimate correspondences across frames. Automatic white balancing, which cannot be disabled on many commercial cameras including ours, also makes finding correspondences challenging when lighting conditions are changing rapidly. Most CMOS chips employ a rolling shutter that becomes noticeable in high impact motions. Substantial motion in the scene, that may occur, for example, when recording in a forest on a windy day, are also likely to present challenges as they violate the intrinsic assumptions made by SfM. Despite these limitations, however, as we illustrate, our system is capable of capturing everyday motions outdoors for extended periods of time and without noticeable drift.

Occlusion by other body parts can cause errors in motion estimation. In practice, three mitigation strategies are used. First, the cameras are carefully placed on the body to minimize the probabil-



(a) Swinging on a monkey bar

(b) Swinging on a swing



(c) Running on a street

Figure 8: *Three captures outside of the laboratory environment: (a) Swinging on a monkey bar, (b) on a swing and (c) running on a street. Top rows illustrate recordings from a reference camcorder camera of the performance, approximately matched in time to the rendered results below. We are able to reconstruct these motions even though they are quite dynamic.*

ity of self-occlusion from body parts. For instance, the cameras on the thighs and shins are placed looking outward on the right side of the leg. Second, for body parts that are likely to be occluded such as the pelvis, we place multiple cameras. This redundancy allows us to estimate motion even when some cameras experience self-occlusion. Finally, RANSAC provides robustness in the case of minor occlusions.

Our system requires significant computation power compared to other motion capture systems. The bulk of processing time involves SIFT keypoint detection/matching. For a minute of capture, this step may require a day of processing for all cameras[4]. After matching, each sequence requires approximately 10 hours for 10 iterations of absolute and relative camera registration. The final optimization can take up to 4 hours. However, this process is highly parallelizable and a GPU should be very effective in speeding up this computation.

As consumer demand continues to push camera prices lower and quality higher, motion capture using body-mounted cameras may become the setup of choice for outdoor capture. Smaller cameras will reduce the motion of the camera relative to its limb and also permit other attachment technologies. Cameras are already small enough to be embedded invisibly in clothing. City-scale 3D geometrical models are also starting to emerge [Agarwal et al. 2009] as faster structure-from-motion implementations [Frahm et al. 2010] are introduced. Such large scale models can be directly utilized in our system to contextualize long-term motions and compose motions of multiple people in a single geometrically coherent environment.

[4]Our cameras produce approximately 1000 SIFT matches and approximately 300 inliers per image.

Acknowledgements

We would like to thank Takeo Kanade, Irfan Essa, and Srinivasa Narasimhan for useful discussions on this project. We would also like to thank Moshe Mahler, Valeria Reznitskaya, and Matthew Kaemmerer for their help in modeling and rendering, and Justin Macey for his help in recording the motion capture data.

References

AGARWAL, S., SNAVELY, N., SIMON, I., SEITZ, S. M., AND SZELISKI, R. 2009. Building Rome in a day. In *Proc. International Conference on Computer Vision*, 72–79.

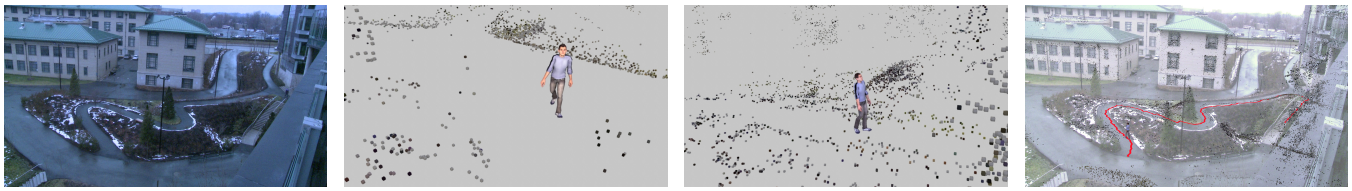


Figure 9: *Subject walking along a long winding path. Left: a photo of the scene, middle two: the reconstructed walking motion and sparse 3D structure, and right: the photo is manually superimposed to the reconstructed scene. The red curve represents the trajectory of the subject.*

BALLAN, L., PUWEIN, J., BROSTOW, G., AND POLLETEYS, M. 2010. Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Transactions on Graphics 29*, 4.

CHEUNG, G. K., BAKER, S., AND KANADE, T. 2003. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 77–84.

CORAZZA, S., MÜNDERMANN, L., CHAUDHARI, A., DEMATTIO, T., COBELLI, C., AND ANDRIACCHI, T. 2006. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. *Annals of Biomedical Engineering 34*, 6, 1019–1029.

CORAZZA, S., GAMBARETTO, E., MÜNDERMANN, L., AND ANDRIACCHI, T. 2010. Automatic generation of a subject-specific model for accurate markerless motion capture and biomechanical applications. *IEEE Transactions on Biomedical Engineering 57*, 4, 806–812.

DAVISON, A., REID, I., MOLTON, N., AND STASSE, O. 2007. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 6, 1052–1067.

DEUTSCHER, J., AND REID, I. 2005. Articulated body motion capture by stochastic search. *International Journal of Computer Vision 61*, 2, 185–205.

DEVERNAY, F., AND FAUGERAS, O. 2000. Straight lines have to be straight. *Machine Vision and Applications 13*, 1, 14–24.

DUNCAN, J. 2010. Avatar. *Cinefex 120* (January), 68–146.

FISCHLER, M., AND BOLLES, R. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6, 381–395.

FRAHM, J.-M., GEORGEL, P., GALLUP, D., JOHNSON, T., RAGURAM, R., WU, C., JEN, Y.-H., DUNN, E., CLIPP, B., LAZEBNIK, S., AND POLLEFEYS, M. 2010. Building Rome on a cloudless day. In *Proc. European Conference on Computer Vision*, 368–381.

HARTLEY, R. I., AND ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

HASLER, N., ROSENHAHN, B., THORMÄHLEN, T., WAND, M., GALL, J., AND SEIDEL, H.-P. 2009. Markerless motion capture with unsynchronized moving cameras. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 224–231.

KELLY, P., CONAIRE, C. Ó., AND O'CONNOR, N. E. 2010. Human motion reconstruction using wearable accelerometers. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation (Poster)*.

KLEIN, G., AND MURRAY, D. 2007. Parallel tracking and mapping for small AR workspaces. In *Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality*, 225–234.

LEPETIT, V., MORENO-NOGUER, F., AND FUA, P. 2009. EP*n*P: An accurate O(n) solution to the P*n*P problem. *International Journal of Computer Vision 81*, 2, 155–166.

LOURAKIS, M. A., AND ARGYROS, A. 2009. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software 36*, 1, 1–30.

LOWE, D. 2004. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision 60*, 2, 91–110.

MOESLUND, T. B., HILTON, A., AND KRÜGER, V. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding 104*, 90–126.

MUJA, M., AND LOWE, D. G. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In *Proc. International Conference on Computer Vision Theory and Application*, 331–340.

NÍSTER, D., NARODITSKY, O., AND BERGEN, J. 2006. Visual odometry for ground vehicle applications. *Journal of Field Robotics 23*, 1, 3–20.

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

OSKIPER, T., ZHU, Z., SAMARASEKERA, S., AND KUMAR, R. 2007. Visual odometry system using multiple stereo cameras and inertial measurement unit. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

POLLEFEYS, M., GOOL, L. V., VERGAUWEN, M., VERBIEST, F., CORNELIS, K., TOPS, J., AND KOCH, R. 2004. Visual modeling with a hand-held camera. *International Journal of Computer Vision 59*, 3, 207–232.

RASKAR, R., NII, H., DE DECKER, B., HASHIMOTO, Y., SUMMET, J., MOORE, D., ZHAO, Y., WESTHUES, J., DIETZ, P., INAMI, M., NAYAR, S., BARNWELL, J., NOLAND, M., BEKAERT, P., BRANZOI, V., AND BRUNS, E. 2007. Prakash: Lighting-aware motion capture using photosensing markers and multiplexed illuminators. *ACM Transactions on Graphics 26*, 3.

SCHWARZ, L. A., MATEUS, D., AND NAVAB, N. 2010. Multiple-activity human body tracking in unconstrained environments. In *Proc. International Conference on Articulated Motion and Deformable Objects*, 192–202.

SLYPER, R., AND HODGINS, J. K. 2008. Action capture with accelerometers. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation*.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics 25*, 3, 835–846.

TAUTGES, J., ZINKE, A., KRÜGER, B., BAUMANN, J., WEBER, A., HELTEN, T., MÜLLER, M., SEIDEL, H.-P., AND EBERHARDT, B. 2011. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics 30*, 3.

VLASIC, D., ADELSBERGER, R., VANNUCCI, G., BARNWELL, J., GROSS, M., MATUSIK, W., AND POPOVIĆ, J. 2007. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics 26*, 3, 35.

WELCH, G., AND FOXLIN, E. 2002. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications 22*, 6, 24–38.

WELCH, G., BISHOP, G., VICCI, L., BRUMBACK, S., KELLER, K., AND COLUCCI, D. 1999. The HiBall tracker: High-performance wide-area tracking for virtual and augmented environments. In *Proc. ACM Symposium on Virtual Reality Software and Technology*, 1–10.

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

XIE, L., KUMAR, M., CAO, Y., GRACANIN, D., AND QUEK, F. 2008. Data-driven motion estimation with low-cost sensors. In *Proc. International Conference on Visual Information Engineering*.

ZHANG, Z., WU, Z., CHEN, J., AND WU, J.-K. 2009. Ubiquitous human body motion capture using micro-sensors. In *Proc. IEEE International Conference on Pervasive Computing and Communications*.

ZHU, Z., OSKIPER, T., SAMARASEKERA, S., SAWHNEY, H., AND KUMAR, R. 2007. Ten-fold improvement in visual odometry using landmark matching. In *Proc. International Conference on Computer Vision*.

ZHU, Z., OSKIPER, T., SAMARASEKERA, S., KUMAR, R., AND SAWHNEY, H. 2008. Real-time global localization with a pre-built visual landmark database. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

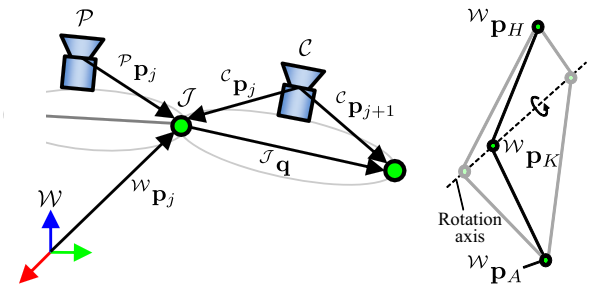
Appendix: Estimating Skeleton from Range-of-Motion

Joints are a point that connects the parent and child limbs, and these limbs are associated to the parent camera $\mathcal{P}$ and the child camera $\mathcal{C}$. While the joint positions in the world coordinate system, ${}^{\mathcal{W}}\mathbf{p}_j$, change over time, the joint positions in the parent and child camera coordinate systems, ${}^{\mathcal{P}}\mathbf{p}_j$ and ${}^{\mathcal{C}}\mathbf{p}_j$, are constant [O'Brien et al. 2000] (Figure 10(a)):

$$ {}^{\mathcal{W}}\mathbf{p}_j(t) = {}^{\mathcal{W}}\mathbf{T}_{\mathcal{P}}(t)\,{}^{\mathcal{P}}\mathbf{p}_j = {}^{\mathcal{W}}\mathbf{T}_{\mathcal{C}}(t)\,{}^{\mathcal{C}}\mathbf{p}_j, \tag{12} $$

where ${}^{\mathcal{W}}\mathbf{T}_{\mathcal{P}}$ and ${}^{\mathcal{W}}\mathbf{T}_{\mathcal{C}}$ are 4×4 Euclidean transformation matrices from the parent and child camera coordinate systems to the world coordinate system, respectively. Equation (12) follows that

$$ \begin{aligned} {}^{\mathcal{P}}\mathbf{p}_j &= {}^{\mathcal{W}}\mathbf{T}_{\mathcal{P}}(t)^{-1}\,{}^{\mathcal{W}}\mathbf{T}_{\mathcal{C}}(t)\,{}^{\mathcal{C}}\mathbf{p}_j \\ &= {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t)\,{}^{\mathcal{C}}\mathbf{p}_j. \end{aligned} \tag{13} $$



(a) Joint and associated cameras (b) Knee joint

Figure 10: *(a) The skeleton is parameterized by parent and child camera poses, $\mathcal{P}$ and $\mathcal{C}$ using local coordinate vectors, ${}^{\mathcal{C}}\mathbf{p}_j$ and ${}^{\mathcal{C}}\mathbf{p}_{j+1}$. (b) One-DOF joints produce a family of solutions for a joint position that lie on the axis of rotation. By assuming the rest pose is a fully extended extremity, where both limbs coincident at the joint are co-linear, joint position can be regularized.*

Thus, collecting Equation (13) for the j-th joint across time provides the homogeneous equation for ${}^{\mathcal{C}}\mathbf{p}_j$,

$$ \begin{bmatrix} {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_1) - {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_2) \\ {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_1) - {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_3) \\ \vdots \\ {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_1) - {}^{\mathcal{P}}\mathbf{T}_{\mathcal{C}}(t_T) \end{bmatrix} {}^{\mathcal{C}}\mathbf{p}_j = \Delta\mathbf{T}\,{}^{\mathcal{C}}\mathbf{p}_j = \mathbf{0}. \tag{14} $$

For two or three-DOF ball joints, the right null vector of $\Delta\mathbf{T}$ obtained with singular value decomposition (SVD) is a solution of ${}^{\mathcal{C}}\mathbf{p}_j$. ${}^{\mathcal{P}}\mathbf{p}_j$ can be also computed in a similar way.

To obtain the skeleton for the whole body, the (j+1)-th joint position from the parent joint in the corresponding camera coordinate system, ${}^{\mathcal{J}}\mathbf{q}$, is computed for each limb as

$$ {}^{\mathcal{J}}\mathbf{q} = \begin{bmatrix} {}^{\mathcal{J}}\widetilde{\mathbf{p}}_{j+1} - {}^{\mathcal{J}}\widetilde{\mathbf{p}}_j \\ 1 \end{bmatrix}, \tag{15} $$

where $\widetilde{\mathbf{p}}$ is an inhomogeneous coordinate of $\mathbf{p}$, and $\mathcal{J}$ is the joint coordinate system.

Additional Constraint on Knees: Knees are one-DOF hinge joints, and Equation (14) becomes an undetermined system: two null vectors can be obtained from $\Delta\mathbf{T}$, and the knee joint position in the thigh camera coordinate system, ${}^{\mathcal{C}_T}\mathbf{p}_K$, is a linear combination of the null vectors (Figure 10(b)):

$$ {}^{\mathcal{C}_T}\mathbf{p}_K = \mathbf{V}_K\mathbf{c}, \tag{16} $$

where $\mathbf{V}_K$ is a matrix consisting of the two null vectors of $\Delta\mathbf{T}$ and $\mathbf{c}$ is a 2D coefficient vector for the null vectors. To determine $\mathbf{c}$, we consider the collinearity constraint caused by straight knees in the rest pose. This collinearity constraint is represented as

$$ \left[\ {}^{\mathcal{W}}\mathbf{p}_H - {}^{\mathcal{W}}\mathbf{p}_A\ \right]_\times \left({}^{\mathcal{W}}\mathbf{p}_K - {}^{\mathcal{W}}\mathbf{p}_A\right) = \mathbf{0}, \tag{17} $$

where ${}^{\mathcal{W}}\mathbf{p}_H$ and ${}^{\mathcal{W}}\mathbf{p}_A$ are the hip and ankle joint positions, and $[\cdot]_\times$ is the skew-symmetric representation for vector cross product. The collinearlity constraint enables a unique solution of the knee joint positions.