To appear in ACM TOG (SIGGRAPH Asia 2016).

EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras

Helge Rhodin[1] Christian Richardt[1, 2, 3] Dan Casas[1] Eldar Insafutdinov[1]
Mohammad Shafiei[1] Hans-Peter Seidel[1] Bernt Schiele[1] Christian Theobalt[1]

[1]Max Planck Institute for Informatics [2]Intel Visual Computing Institute [3]University of Bath

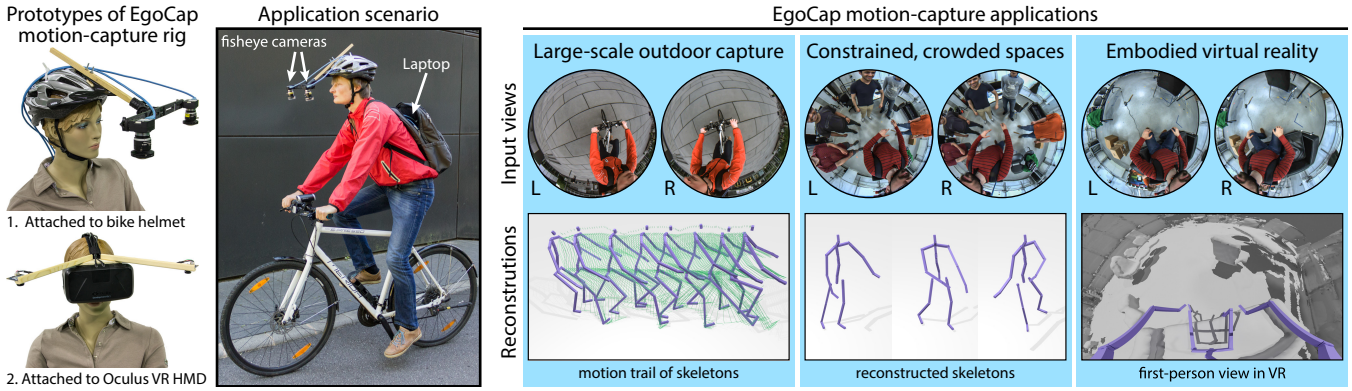


Figure 1: *We propose a marker-less optical motion-capture approach that only uses two head-mounted fisheye cameras (see rigs on the left). Our approach enables three new application scenarios: (1) capturing human motions in outdoor environments of virtually unlimited size, (2) capturing motions in space-constrained environments, e.g. during social interactions, and (3) rendering the reconstruction of one's real body in virtual reality for embodied immersion.*

Abstract

Marker-based and marker-less optical skeletal motion-capture methods use an *outside-in* arrangement of cameras placed around a scene, with viewpoints converging on the center. They often create discomfort with marker suits, and their recording volume is severely restricted and often constrained to indoor scenes with controlled backgrounds. Alternative suit-based systems use several inertial measurement units or an exoskeleton to capture motion with an *inside-in* setup, i.e. without external sensors. This makes capture independent of a confined volume, but requires substantial, often constraining, and hard to set up body instrumentation. Therefore, we propose a new method for real-time, marker-less, and egocentric motion capture: estimating the full-body skeleton pose from a lightweight stereo pair of fisheye cameras attached to a helmet or virtual reality headset – an *optical inside-in* method, so to speak. This allows full-body motion capture in general indoor and outdoor scenes, including crowded scenes with many people nearby, which enables reconstruction in larger-scale activities. Our approach combines the strength of a new generative pose estimation framework for fisheye views with a ConvNet-based body-part detector trained on a large new dataset. It is particularly useful in virtual reality to freely roam and interact, while seeing the fully motion-captured virtual body.

Keywords: Motion capture, first-person vision, markerless, optical, inside-in, crowded scenes, large-scale

Concepts: •**Computing methodologies** → **Motion capture;**

1 Introduction

Traditional optical skeletal motion-capture methods – both marker-based and marker-less – use several cameras typically placed around a scene in an *outside-in* arrangement, with camera views approximately converging in the center of a confined recording volume. This greatly constrains the spatial extent of motions that can be recorded; simply enlarging the recording volume by using more cameras, for instance to capture an athlete, is not scalable. Outside-in arrangements also constrain the type of scene that can be recorded, even if it fits into a confined space. If a recording location is too small, cameras can often not be placed sufficiently far away. In other cases, a scene may be cluttered with objects or furniture, or other dynamic scene elements, such as people in close interaction, may obstruct a motion-captured person in the scene or create unwanted dynamics in the background. In such cases, even state-of-the-art outside-in marker-less optical methods that succeed with just a few cameras, and are designed for less controlled and outdoor scenes [Elhayek et al. 2015], quickly fail. Scenes with dense social interaction were previously captured with outside-in camera arrays of a few hundred sensors [Joo et al. 2015], a very complex and difficult to scale setup.

These strong constraints on recording volume and scene density prevent the use of optical motion capture in the majority of real-world scenes. This problem can partly be bypassed with *inside-in* motion-capture methods that use body-worn sensors exclusively [Menache 2010], such as the Xsens MVN inertial measurement unit suit. However, the special suit and cabling are obstructive and require tedious calibration. Shiratori et al. [2011] propose to wear 16 cameras placed on body parts facing *inside-out*, and capture the skeletal motion through structure-from-motion relative to the environment. This clever solution requires instrumentation, calibration and a static background, but allows free roaming. This design was inspirational for our egocentric approach.

We propose EgoCap: an egocentric motion-capture approach that estimates full-body pose from a pair of optical cameras carried by lightweight headgear (see Figure 1). The body-worn cameras are oriented such that their field of view covers the user's body entirely, forming an arrangement that is independent of external sensors – an *optical inside-in* method, if you will. We show that our optical full-body approach overcomes many limitations of existing outside-in, inside-out and IMU-based inside-in methods. It reduces the setup effort, enables free roaming, and minimizes body instrumentation. EgoCap decouples the estimation of local body pose with respect to the headgear cameras and global headgear position, which we infer by inside-out structure-from-motion on the scene background.

Our first contribution is a new egocentric inside-in sensor rig with only two head-mounted, downward-facing commodity video cameras with fisheye lenses (see Figure 1). While head-mounted cameras might pose a problem with respect to social acceptance and ergonomics in some scenarios, performances have not been hindered during our recordings and VR tests. The rig can be attached to a helmet or a head-mounted VR display, and, hence, requires less instrumentation and calibration than other body-worn systems. The stereo fisheye optics keep the whole body in view in all poses, despite the cameras' proximity to the body. We prefer conventional video cameras over IR-based RGB-D cameras, which were for example used for egocentric hand tracking [Sridhar et al. 2015], as video cameras work indoors and outdoors, have lower energy consumption and are easily fitted with the required fisheye optics.

Our second contribution is a new marker-less motion capture algorithm tailored to the strongly distorted egocentric fisheye views. It combines a generative model-based skeletal pose estimation approach (Section 4) with evidence from a trained ConvNet-based body part detector (Section 4.3). The approach features an analytically differentiable objective energy that can be minimized efficiently, is designed to work with unsegmented frames and general backgrounds, succeeds even on poses exhibiting notable self-occlusions (e.g. when walking), as the part detector predicts occluded parts, and enables recovery from tracking errors after severe occlusions.

Our third contribution is a new approach for automatically creating body part detection training datasets. We record test subjects in front of green screen with an existing outside-in marker-less motion capture system to get ground-truth skeletal poses, which are reprojected into the simultaneously recorded head-mounted fisheye views to get 2D body part annotations. We augment the training images by replacing the green screen with random background images, and vary the appearance in terms of color and shading by intrinsic recoloring [Meka et al. 2016]. With this technique, we annotate a total of 100,000 egocentric images of eight people in different clothing (Section 4.3.1), with 75,000 images from six people used for training. We publish the dataset for research purposes [EgoCap 2016].

We designed and extensively tested two system prototypes featuring (1) cameras fitted to a bike helmet, and (2) small cameras attached to an Oculus Rift headset. We show reliable egocentric motion capture, both off-line and in real time. The egocentric tracking meets the accuracy of outside-in approaches using 2–3 cameras; additional advances are necessary to match the accuracy of many-camera systems. In our egocentric setup, reconstructing the lower body is more challenging due to its larger distance and frequent occlusions, and is less accurate compared to the upper body in our experiments. Nevertheless, we succeed in scenes that are challenging for outside-in approaches, such as close interaction with many people, as well outdoor and indoor scenes in cluttered environments with frequent occlusions, for example when working in a kitchen or at a desk. We also show successful capturing in large volumes, for example of the skeletal motion of a cyclist. The lightweight Oculus Rift gear is designed for egocentric motion capture for virtual reality, where the user can move in the real world to roam and interact in a virtual environment seen through a head-mounted display, while perceiving increased immersion thanks to the rendering of the motion-captured body, which is not obtained with current HMD head pose tracking.

2 Related Work

Suit-based Motion Capture Marker-based optical systems use a suit with passive retro-reflective spheres (e.g. Vicon) or active LEDs (e.g. PhaseSpace). Skeleton motion is reconstructed from observed marker positions in multiple cameras (usually 10 or more) in an outside-in arrangement, producing highly accurate sparse motion data, even of soft tissue [Park and Hodgins 2008, Loper et al. 2014], but the external cameras severely restrict the recording volume. For character animation purposes, where motions are restricted, use of motion sub-spaces can reduce requirements to six markers and two cameras [Chai and Hodgins 2005], or a single foot pressure-sensor pad [Yin and Pai 2003], which greatly improves usability. For hand tracking, a color glove and one camera [Wang and Popović 2009] is highly practical. Inertial measurement units (IMUs) fitted to a suit (e.g. Xsens MVN) allow free roaming and high reliability in cluttered scenes by inside-in motion capture, i.e. without requiring external sensors [Tautges et al. 2011]. Combinations with ultrasonic distance sensors [Vlasic et al. 2007], video input [Pons-Moll et al. 2010, 2011], and pressure plates [Ha et al. 2011] suppress the drift inherent to IMU measurements and reduce the number of required IMUs. Besides drift, the instrumentation with IMU sensors is the largest drawback, causing long setup times and intrusion. Exoskeleton suits (e.g. METAmotion Gypsy) avoid drift, but require more cumbersome instrumentation. Turning the standard outside-in capturing approach on its head, Shiratori et al. [2011] attach 16 cameras to body segments in an inside-out configuration, and estimate skeletal motion from the position and orientation of each camera as computed with structure-from-motion. This clever solution – which was inspirational for our egocentric approach – allows free roaming although it requires instrumentation and a static background.

Marker-less Motion Capture Recent years have seen great advances in marker-less optical motion-capture algorithms that track full-body skeletal motions, reaching and outperforming the reconstruction quality of suit- and marker-based approaches [Bregler and Malik 1998, Theobalt et al. 2010, Moeslund et al. 2011, Holte et al. 2012]. Marker-less approaches also typically use an outside-in camera setup, and were traditionally limited to controlled studio environments, or scenes with static, easy-to-segment background, using 8 or more cameras [e.g. Urtasun et al. 2006, Gall et al. 2010, Sigal et al. 2010, 2012, Stoll et al. 2011]. Recent work is moving towards less controlled environments and outdoor scenes, also using fewer cameras [Amin et al. 2009, Burenius et al. 2013, Elhayek et al. 2015, Rhodin et al. 2015], but still in an outside-in configuration. These approaches are well-suited for static studio setups, but share the limitation of constrained recording volumes, and reach their limits in dense, crowded scenes. Joo et al. [2015] use a camera dome with 480 outside-in cameras for motion capture of closely interacting people, but domes do not scale to larger natural scenes.

Motion Capture with Depth Sensors 3D pose estimation is highly accurate and reliable when using multiple RGB-D cameras [Zhang et al. 2014], and even feasible from a single RGB-D camera in real time [e.g. Shotton et al. 2011, Baak et al. 2011, Wei et al. 2012]. However, many active IR-based depth cameras are unsuitable for outdoor capture, have high energy consumption, and equipping them with fisheye optics needed for our camera placement is hard.

Egocentric Motion Capture In the past, egocentric inside-in camera placements were used for tracking or model learning of certain parts of the body, for example of the face with a helmet-mounted

camera or rig [Jones et al. 2011, Wang et al. 2016], of fingers from a wrist-worn camera [Kim et al. 2012], or of eyes and eye gaze from cameras in a head-mounted rig [Sugano and Bulling 2015]. Rogez et al. [2014] and Sridhar et al. [2015] track articulated hand motion from body- or chest-worn RGB-D cameras. Using a body-worn depth camera, Yonemoto et al. [2015] extrapolate arm and torso poses from arm-only RGB-D footage. Jiang and Grauman [2016] attempted full-body pose estimation from a chest-worn camera view by analyzing the scene, but without observing the user directly and at very restricted accuracy. Articulated full-body motion capture with a lightweight head-mounted camera pair was not yet attempted.

First-person Vision A complementary research branch analyses the environment from first-person, i.e. body-worn outward-facing cameras, for activity recognition [e.g. Fathi et al. 2011, Kitani et al. 2011, Ohnishi et al. 2016, Ma et al. 2016], for learning engagement and saliency patterns of users when interacting with the real world [e.g. Park et al. 2012, Su and Grauman 2016], and for understanding the utility of surrounding objects [Rhinehart and Kitani 2016]. Articulated full-body tracking, or even only arm tracking, is not their goal, but synergies of both fields appear promising.

2D and 3D Pose Detection Traditionally, 2D human pose estimation from monocular images is a two-stage process where coherent body pose is inferred from local image evidence [Yang and Ramanan 2013, Johnson and Everingham 2011]. Convolutional networks (ConvNets) brought a major leap in performance [Chen and Yuille 2014, Jain et al. 2014, 2015, Tompson et al. 2014, Toshev and Szegedy 2014] and recent models demonstrated that end-to-end prediction is possible due to the large receptive fields capturing the complete pose context [Pishchulin et al. 2016]. Pfister et al. [2015] and Wei et al. [2016] allow for increased depth and learning of spatial dependencies between body parts by layering multiple ConvNets. We adopt the network architecture of Insafutdinov et al. [2016], which builds on the recent success of residual networks [He et al. 2016, Newell et al. 2016], which further facilitate an increase in network depth. Recently, direct 3D pose estimation has emerged by lifting 2D poses to 3D [Yasin et al. 2016], using mid-level posebit descriptors [Pons-Moll et al. 2014], and motion compensation in videos [Tekin et al. 2016], but estimates are still coarse. Existing detection methods use simplified body models with few body parts to reduce the enormous cost of creating sufficiently large, annotated training datasets, do not generalize to new camera geometry and viewpoints, such as egocentric views, and results usually exhibit jitter over time.

3 Egocentric Camera Design

We designed a mobile egocentric camera setup to enable human motion capture within a virtually unlimited recording volume. We attach two fisheye cameras rigidly to a helmet or VR headset, such that their field of view captures the user's full body, see Figure 2. The wide field of view allows to observe interactions in front and beside the user, irrespective of their global motion and head orientation, and without requiring additional sensors or suits. The stereo setup ensures that most actions are observed by at least one camera, despite substantial self-occlusions of arms, torso and legs in such an egocentric setup. A baseline of 30–40 cm proved to be best in our experiments. The impact of the headgear on the user's motion is limited as it is lightweight: our prototype camera rig for VR headsets (see Figure 1, bottom left) only adds about 65 grams of weight.

4 Egocentric Full-Body Motion Capture

Our egocentric setup separates human motion capture into two subproblems: (1) local skeleton pose estimation with respect to the camera rig, and (2) global rig pose estimation relative to the environment. Global pose is estimated with existing structure-from-motion techniques, see Section 6.3. We formulate skeletal pose estimation as an analysis-by-synthesis-style optimization problem in the pose parameters $\mathbf{p}^t$, that maximizes the alignment of a projected 3D human body model (Section 4.1) with the human in the left $\mathcal{I}^t_{\text{left}}$ and the right $\mathcal{I}^t_{\text{right}}$ stereo fisheye views, at each video time step t. We use a hybrid alignment energy combining evidence from a generative image-formation model, as well as from a discriminative detection approach. Our generative ray-casting-based image formation model is inspired by light transport in volumetric translucent media, and enables us to formulate a color-based alignment term in $\mathbf{p}^t$ that is analytically differentiable and features an analytically differentiable formulation of 3D visibility (Section 4.2). This model facilitates generative pose estimation with only two cameras, and we adapt it to the strongly distorted fisheye views. Our energy also employs constraints from one-shot joint-location predictions in the form of $E_{\text{detection}}$. These predictions are found with a new ConvNet-based 2D joint detector for head-mounted fisheye views, which is learned from a large corpus of annotated training data, and which generalizes to different users and cluttered scenes (Section 4.3). The combined energy that we optimize takes the following form:

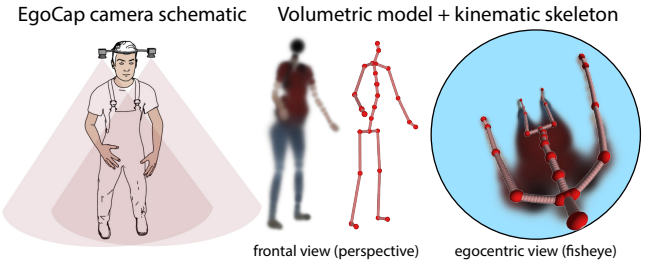


Figure 2: *Schematic of EgoCap, our egocentric motion-capture rig (left), visualization of the corresponding volumetric body model and kinematic skeleton (center), and the egocentric view of both in our head-mounted fisheye cameras (right).*

$$E(\mathbf{p}^t) = E_{\text{color}}(\mathbf{p}^t) + E_{\text{detection}}(\mathbf{p}^t) + E_{\text{pose}}(\mathbf{p}^t) + E_{\text{smooth}}(\mathbf{p}^t). \quad (1)$$

Here, $E_{\text{pose}}(\mathbf{p}^t)$ is a regularizer that penalizes violations of anatomical joint-angle limits as well as poses deviating strongly from the rest pose ($\mathbf{p} = \mathbf{0}$):

$$E_{\text{pose}}(\mathbf{p}^t) = \lambda_{\text{limit}} \cdot \Big(\max(0, \mathbf{p}^t - \mathbf{l}_{\text{upper}})^2 + \max(0, \mathbf{l}_{\text{lower}} - \mathbf{p}^t)^2 \Big) + \lambda_{\text{pose}} \cdot \text{huber}(\mathbf{p}^t), \quad (2)$$

where $\mathbf{l}_{\text{lower}}$ and $\mathbf{l}_{\text{upper}}$ are lower and upper joint-angle limits, and $\text{huber}(x) = \sqrt{1+x^2} - 1$ is the Pseudo-Huber loss function. $E_{\text{smooth}}(\mathbf{p}^t)$ is a temporal smoothness term:

$$E_{\text{smooth}}(\mathbf{p}^t) = \lambda_{\text{smooth}} \cdot \text{huber}(\mathbf{p}^{t-1} + \zeta(\mathbf{p}^{t-1} - \mathbf{p}^{t-2}) - \mathbf{p}^t), \quad (3)$$

where $\zeta = 0.25$ is a damping factor. The total energy in Equation 1 is optimized for every frame, as described in Section 4.4. In the following, we describe the generative and discriminative terms in more detail, while omitting the temporal dependency t in the notation for better readability.

We use weights $\lambda_{\text{pose}} = 10^{-4}$, $\lambda_{\text{limit}} = 0.1$ and $\lambda_{\text{smooth}} = 0.1$.

4.1 Body Model

We model the 3D body shape and pose of humans in 3D using the approach proposed by Rhodin et al. [2015], which represents

the body volumetrically as a set of $N_q = 91$ isotropic Gaussian density functions distributed in 3D space. Each Gaussian G_q is parametrized by its standard deviation σ_q, location $\boldsymbol{\mu}_q$ in 3D space, density c_q and color $\mathbf{a}_q$, which define the Gaussian shape parameters. The combined density field of the Gaussians, $\sum_q c_q G_q$, smoothly describes the volumetric occupancy of the human in 3D space, see Figure 2. Each Gaussian is rigidly attached to one of the bones of an articulated skeleton with 17 joints, whose pose is parameterized by 37 twist pose parameters [Murray et al. 1994].

Shape and skeleton bone lengths need to be personalized to the tracked user prior to capturing. Commercial systems often use a dedicated initialization sequence at the start. Research papers on marker-less motion capture often treat initialization as a separate problem, and initialize models manually, which we could also do. However, we propose a much more automated initialization procedure to reduce setup time and effort. To this end, we adapt the approach of Rhodin et al. [2016], who personalize a 3D parametric human shape model of Gaussian density and skeleton dimensions by fitting it to multi-view images using a volumetric contour alignment energy. We adapt this to our stereo fisheye setting. In our egocentric setup 3–4 different user poses, showing the bending of knees, elbows and wrists without any occlusion, were sufficient for automatic shape and skeleton personalization, and only the automatically inferred Gaussian colors are manually corrected on body parts viewed at acute angles.

4.2 Egocentric Volumetric Ray-Casting Model

For color-based model-to-image similarity, we use the ray-casting image formation model of the previously described volumetric body model [Rhodin et al. 2015]. We first describe image formation assuming a standard pinhole model, as in Rhodin et al., and then describe how we modify it for fisheye views. A ray is cast from the camera center $\mathbf{o}$ in direction $\mathbf{n}$ of an image pixel. The visibility of a particular 3D Gaussian G_q along the ray $(\mathbf{o} + s\mathbf{n})$ is computed via

$$\mathcal{V}_q(\mathbf{o},\mathbf{n},\mathbf{p}) = \int_0^\infty \exp\left(-\int_0^s \sum_i G_i(\mathbf{o}+t\mathbf{n})\,\mathrm{d}t\right) G_q(\mathbf{o}+s\mathbf{n})\,\mathrm{d}s. \quad (4)$$

This formulation of visibility and color of a 3D Gaussian from the camera view is based on a model of light transport in heterogeneous translucent media [Cerezo et al. 2005]. $\mathcal{V}_q$ is the fraction of light along the ray that is absorbed by Gaussian G_q. We use this image-formation model in an energy term that computes the agreement of model and observation by summing the visibility-weighted color dissimilarity $d(\cdot,\cdot)$, which we describe in Appendix A, between image pixel color $\mathcal{I}(u,v)$ and the Gaussian's color $\mathbf{a}_q$:

$$E_{\text{color}}(\mathbf{p},\mathcal{I}) = \sum_{(u,v)} \sum_q d(\mathcal{I}(u,v),\mathbf{a}_q)\mathcal{V}_q(\mathbf{o},\mathbf{n}(u,v),\mathbf{p}). \quad (5)$$

Note that this formulation has several key advantages over previous generative models for image-based pose estimation. It enables analytic derivatives of the pose energy, including a smooth analytically differentiable visibility model everywhere in pose space. This makes it perform well with only a few camera views. Previous methods often used fitting energies that are non-smooth or even lacking a closed-form formulation, requiring approximate recomputation of visibility (e.g. depth testing) inside an iterative optimization loop. Rhodin et al.'s formulation forms a good starting point for our egocentric tracking setting, as non-stationary backgrounds and occlusions are handled well. However, it applies only to static cameras, does not support the distortion of fisheye lenses, and it does not run in real time.

4.2.1 Egocentric Ray-Casting Model

In our egocentric camera rig, the cameras move rigidly with the user's head. In contrast to commonly used skeleton configurations, where the hip is taken as the root joint, our skeleton hierarchy is rooted at the head. Like a puppet, the lower body parts are then relative to the head motion, see Figure 2. This formulation factors out the user's global motion, which can be estimated independently, see Section 6.3, and reduces the dimensionality of the pose estimation by 6 degrees of freedom. By attaching the cameras to the skeleton root, the movable cameras are reduced to a static camera formulation such that Equation 4 applies without modification.

Simply undistorting the fisheye images before optimization is impractical as resolution at the image center reduces and pinhole cameras cannot capture fields of view approaching 180 degrees – their image planes would need to be infinitely large. To apply the ray-casting formulation described in the previous section to our egocentric motion-capture rig, with its 180° field of view, we replace the original pinhole camera model with the omnidirectional camera model of Scaramuzza et al. [2006]. The ray direction $\mathbf{n}(u,v)$ of a pixel (u,v) is then given by $\mathbf{n}(u,v) = [u, v, f(\rho)]^\top$, where f is a polynomial of the distance ρ of (u,v) to the estimated image center. We combine the energy terms for the two cameras (Equation 5) in our egocentric camera rig using

$$E_{\text{color}}(\mathbf{p}) = E_{\text{color}}(\mathbf{p},\mathcal{I}_{\text{left}}) + E_{\text{color}}(\mathbf{p},\mathcal{I}_{\text{right}}). \quad (6)$$

These extensions also generalize the contour model of Rhodin et al. [2016] to enable egocentric body model initialization.

4.3 Egocentric Body-Part Detection

We combine the generative model-based alignment from the previous section with evidence from the discriminative joint-location detector of Insafutdinov et al. [2016], trained on annotated egocentric fisheye images. The discriminative component dramatically improves the quality and stability of reconstructed poses, provides efficient recovery from tracking failures, and enables plausible tracking even under notable self-occlusions. To apply Insafutdinov et al.'s body-part detector, which has shown state-of-the-art results on human pose estimation from outside-in RGB images, to the top-down perspective and fisheye distortion of our novel egocentric camera setup, the largest burden is to gather and annotate a training dataset that is sufficiently large and varied, containing tens of thousands of images. As our camera rig is novel, there are no existing public datasets, and we therefore designed a method to automatically annotate real fisheye images by outside-in motion capture and to augment appearance with the help of intrinsic image decomposition.

4.3.1 Dataset Creation

We propose a novel approach for semi-automatically creating large, realistic training datasets for body-part detection that comprise tens of thousands of camera images annotated with the joint locations of a kinematic skeleton and other body parts such as the hands and feet. To avoid the tedious and error-prone manual annotation of locations in thousands of images, as in previous work, we use a state-of-the-art marker-less motion capture system (Captury Studio of The Captury) to estimate the skeleton motion in 3D from eight stationary cameras placed around the scene. We then project the skeleton joints into the fisheye images of our head-mounted camera rig. The projection requires tracking the rigid motion of our head-mounted camera rig relative to the stationary cameras of the motion-capture system, for which we use a large checkerboard rigidly attached to our camera rig (Figure 3). We detect the checkerboard in all stationary cameras in which it is visible, and triangulate the 3D positions of its corners

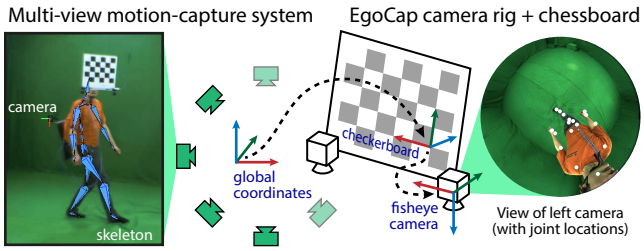


Figure 3: *For database annotation, the skeleton estimated from the multi-view motion capture system (left), is converted from global coordinates (center) into each fisheye camera's coordinate system (right) via the checkerboard.*

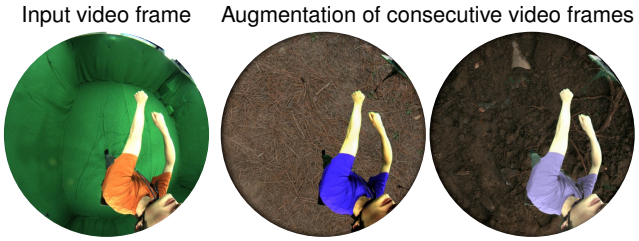


Figure 4: *Illustration of our dataset augmentation using randomized backgrounds, intrinsic recoloring and gamma jittering. Note the varied shirt colors as well as brightness of the trousers and skin, which help prevent overtraining of the ConvNet-based joint detector.*

to estimate the pose and orientation of the camera rig. Using Scaramuzza et al.'s camera distortion model, we then project the 3D joint locations into the fisheye images recorded by our camera rig.

Dataset Augmentation We record video sequences of eight subjects performing various motions in a green-screen studio. For the training set, we replace the background of each video frame, using chroma keying, with a random, floor-related image from Flickr, as our fisheye cameras mostly see the ground below the tracked subject. Please note that training with real backgrounds could give the CNN additional context, but is prone to overfitting to a (necessarily) small set of recorded real backgrounds. In addition, we augment the appearance of subjects by varying the colors of clothing, while preserving shading effects, using intrinsic recoloring [Meka et al. 2016]. This is, to our knowledge, the first application of intrinsic recoloring for augmenting datasets. We also apply a random gamma curve ($\gamma \in [0.5, 2]$) to simulate changing lighting conditions. We furthermore exploit the shared plane of symmetry of our camera rig and the human body to train a single detector on a dataset twice the size by mirroring the images and joint-location annotations of the right-hand camera to match those of the left-hand camera during training, and vice versa during motion capture. Thanks to the augmentation, both background and clothing colors are different for every frame (see Figure 4), which prevents overfitting to the limited variety of the captured appearances. This results in a training set of six subjects and ~75,000 annotated fisheye images. Two additional subjects are captured and prepared for validation purposes.

4.3.2 Detector Learning

Our starting point for learning an egocentric body-part detector for fisheye images is the 101-layer residual network [He et al. 2016] trained by Insafutdinov et al. [2016] on the MPII Human Pose dataset [Andriluka et al. 2014], which contains ~19,000 internet images that were manually annotated in a crowd-sourced effort, and the Leeds Sports Extended dataset [Johnson and Everingham 2011]

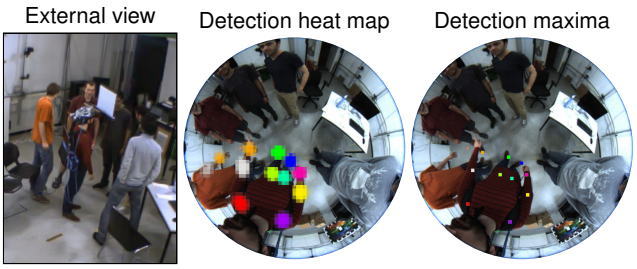


Figure 5: *Color-coded joint-location detections on the* `Crowded` *sequence. For crowded scenes (left), detections can be multi-modal (center). However, the maximum (right) lies on the user. We exclude knee, hand and ankle locations for clearer visualization.*

of 10,000 images. We remove the original prediction layers and replace them with ones that output 18 body-part heat maps[1]. The input video frames are scaled to a resolution of 640×512 pixels, the predicted heat maps are of 8× coarser resolution. We then fine-tune the ConvNet on our fisheye dataset for 220,000 iterations with a learning rate of 0.002, and drop it to 0.0002 for 20,000 additional iterations. The number of training iterations is chosen based on performance on the validation set. We randomly scale images during training by up to ±15 to be more robust to variations in user size. Figure 5 (center) visualizes the computed heat maps for selected body parts. We demonstrate generalization capability to a large variety of backgrounds, changing illumination and clothing colors in Section 5.3.

4.3.3 Body-Part Detection Energy

Inspired by Elhayek et al. [2015], who exploit detections in outside-in motion capture, we integrate the learned detections, in the form of heat maps as shown in Figure 5, into the objective energy (Equation 1) as a soft constraint. For each detection label, the location with maximum confidence, $(\hat{u}, \hat{v})$, is selected and an associated 3D Gaussian is attached to the corresponding skeleton body part. This association can be thought of as giving a distinct color to each body-part label. The Gaussian is used to compute the spatial agreement of the detection and body-part location in the same way as in the color similarity E_{color}, only the color distance $d(\cdot,\cdot)$ in Equation 5 is replaced with the predicted detection confidence at $(\hat{u}, \hat{v})$. For instance, a light green Gaussian is placed at the right knee and is associated with the light green knee detection heat map at $(\hat{u}, \hat{v})$, then their agreement is maximal when the Gaussian's center projects on $(\hat{u}, \hat{v})$. By this definition, $E_{\text{detection}}$ forms the sum over the detection agreements of all body parts and in both cameras. We weight its influence by $\lambda_{\text{detection}} = 1/3$.

4.4 Real-Time Optimization

Rhodin et al.'s volumetric ray-casting method [2015] models occlusion as a smooth phenomenon by integrating the visibility computations within the objective function instead of applying a depth test once before optimization. While this is beneficial for optimizing disocclusions, it introduces dense pairwise dependencies between all Gaussians: the visibility $\mathcal{V}_q$ (Equation 4) of a single Gaussian can be evaluated in linear time in terms of the number of Gaussians, N_q, but E_{color} – and its gradient with respect to all Gaussians – has quadratic complexity in N_q.

To nevertheless reach real-time performance, we introduce a new par-

[1]We jointly learn heat maps for the head and neck, plus the left and right shoulders, elbows, wrists, hands, hips, knees, ankles and feet.

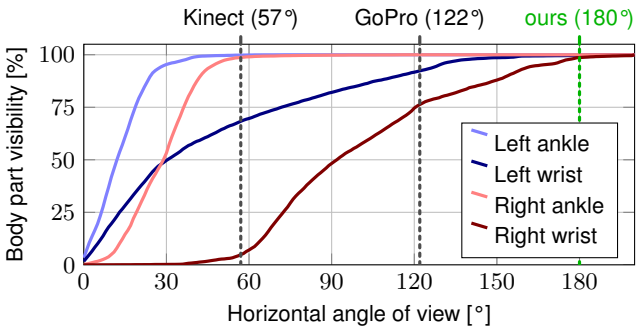


Figure 6: *Visibility of selected body parts for different camera angles of view, for the left-hand camera in our rig over a 5-minute recording. Seeing the right wrist 95 percent of the time requires an angle of view in excess of 160°, which is only practical with fisheye lenses.*

allel stochastic optimization approach. The ray-casting formulation allows a natural parallelization of $E_{\text{detection}}$ and E_{color} terms and their gradient computation across pixels (u, v) and Gaussians G_q. We also introduce a traversal step, which determines the Gaussians that are close to each ray, and excludes distant Gaussians with negligible contribution to the energy. These optimizations lead to significant run-time improvements, particularly when executed on a GPU, but only enable interactive frame rates.

We achieve further reductions in run times by introducing a statistical optimization approach that is tailored to the ray-casting framework. The input image pixels are statistically sampled for each gradient iteration step, as proposed by Blanz and Vetter [1999]. In addition, we sample the volumetric body model by excluding Gaussians from the gradient computation at random, individually for each pixel, which improves the optimization time to 10 fps and more.

5 Evaluation

5.1 Hardware Prototypes

We show the two EgoCap prototypes used in this work in Figure 1 (left). *EgoRig1* consists of two fisheye cameras attached to a standard bike helmet. It is robust and well-suited for capturing outdoor activities and sports. *EgoRig2* builds on a lightweight wooden rig that holds two consumer cameras and is glued to an Oculus VR headset. It weighs only 65 grams and adds minimal discomfort on the user. Both prototypes are equipped with 180° fisheye lenses and record with a resolution of 1280×1024 pixels at 30 Hz. Note that the checkerboard attached to *EgoRig1* in several images is not used for tracking (only used in training and validation dataset recordings).

Body-Part Visibility For egocentric tracking of unconstrained motions, the full 180° field of view is essential for egocentric tracking. We evaluate the visibility of selected body parts from our egocentric rig with different (virtual) field-of-view angles in Figure 6. Only at 180 degrees are almost all body parts captured, otherwise even small motions of the head can cause the hand to leave the recording volume. The limited field of view of existing active depth sensors of 60–80 degrees restricts their applicability to egocentric motion capture in addition to their higher energy consumption and interference with other light sources.

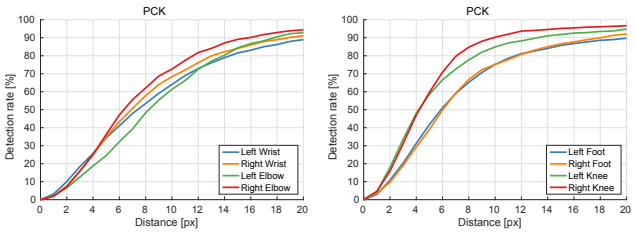
5.2 Runtime

For most tracking results, we use a resolution of 128×128 pixels and 200 gradient-descent iterations. Our CPU implementation runs at ten seconds per frame on a Xeon E5-1620 3.6 GHz, which is similar to run times reported by Rhodin et al. [2015]. Straightforward parallelization on the GPU reduces run times to two seconds per frame. The body-part detector runs on a separate machine, and processes 6 images per second on an Nvidia Titan GPU and a Xeon E5-2643 3.30 GHz.

For some experiments (see Section 6.3), we use a resolution of 120×100 pixels and enable stochastic optimization. Then, purely color-based optimization reaches 10 to 15 fps for 50 gradient iterations (2–3 ms per iteration), i.e. close to real-time performance. Our body-part detector is not optimized for speed and cannot yet run at this frame rate, but its implementation could be optimized for real-time processing, so a real-time end-to-end approach would be feasible without algorithmic changes.

Table 1: *Part detection accuracy in terms of the percentage of correct keypoints (PCK) on the validation dataset* `Validation2D` *of 1000 images, evaluated at 20 pixel threshold for three ConvNets trained with different data augmentation strategies (Section 4.3.1). AUC is area under curve evaluated for all thresholds up to 20 pixels.*

Training dataset setting	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCK	AUC
green-screen background	75.5	46.8	18.8	13.6	17.4	7.2	4.5	22.4	10.0
+ background augmentation	84.7	87.5	90.9	89.1	97.7	94.2	86.4	89.5	56.9
+ intrinsic recoloring	**86.2**	**96.1**	**93.6**	**90.1**	**99.1**	**95.8**	**90.9**	**92.5**	**59.4**



(a) Arm joints (b) Leg joints

Figure 7: *Pose estimation results in terms of percentage of correct keypoints (PCK) for different distance thresholds on* `Validation2D`.

5.3 Body-Part Detections

We first evaluate the learned body-part detectors, irrespective of generative components, using the percentage of correct keypoints (PCK) metric [Sapp and Taskar 2013, Tompson et al. 2014]. We evaluate on a validation set, `Validation2D`, of 1000 images from a 30,000-frame sequence of two subjects that are not part of the training set and wear dissimilar clothing. `Validation2D` is augmented with random backgrounds using the same procedure as for the training set, such that the difficulty of the detection task matches the real-world sequences. We further validated that overfitting to augmentation is minimal, by testing on green-screen background, with equivalent results.

Dataset Augmentations Table 1 presents the evaluation of proposed data augmentation strategies. Background augmentation during training brings a clear improvement. It provides a variety of challenging negative samples for the training of the detector, which is of high importance. Secondly, the performance is further boosted by employing intrinsic video for cloth recoloring, which additionally increases the diversity of training samples. The improvement of about two percent is consistent across all body parts.

Detection Accuracy Figure 7 contains the plots of PCK at different distance thresholds for arms and legs evaluated on sequence `Validation2D`. We achieve high accuracy, with slightly lower detec-

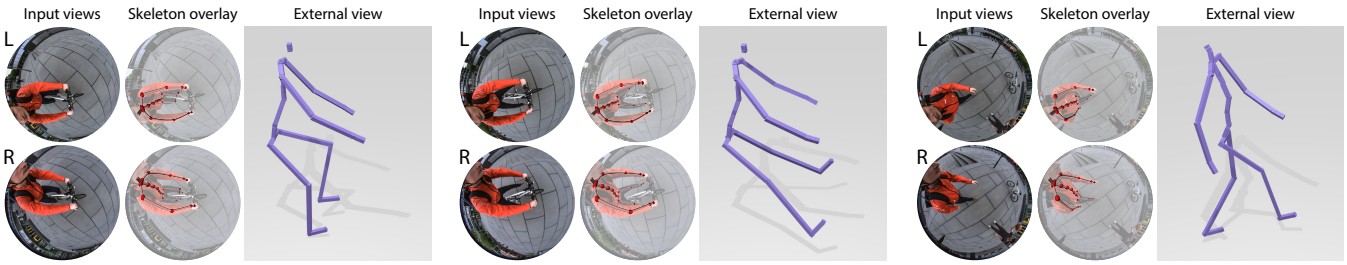


Figure 8: *EgoCap enables outdoor motion capture with virtually unconstrained extent. Full-body pose is accurately estimated for fast* `Biking` *(left and center) and for unconstrained* `Walk` *(right). The model is tailored to handle the present occlusions and strong image distortion.*

tion reliability of terminal limbs (wrists, feet). This can either be due to more articulation or, in case of the feet, due to higher occlusion by knees and their small appearance due to the strong fisheye distortion. The 2D detection accuracy of feet and wrists is comparable, even though feet are further away, and similar pixel error hence translates to larger 3D errors, as evaluated in the next section. We additionally evaluated the training set size. We found that subject variation is important: using only three out of six subjects, the PCK performance dropped by 2.5 percent points. Moreover, using a random subset of 10 of the original database size reduces the PCK by 2 points, i.e. using more than three frames per second is beneficial. Using a 50 subset did not degrade performance, showing that consecutive frames are not crucial for our per-frame model, but could be beneficial for future research, such as for temporal models.

5.4 3D Body Pose Accuracy

Our main objective is to infer 3D human pose from the egocentric views, despite occlusions and strong fisheye image distortions. We quantitatively evaluate the 3D body pose accuracy of our approach on two sequences, `ValidationWalk` and `ValidationGest`. Ground-truth data is obtained with the Captury Studio, a state-of-the-art marker-less commercial multi-view solution with eight video cameras and 1–2 cm accuracy. The two systems are used simultaneously and their relative transformation is estimated with a reference checkerboard, see Figure 3. We experimented with raw green-screen and with randomly replaced background. Error values are estimated as the average Euclidean 3D distance over 17 joints, including all joints with detection labels, except the head. Reconstructions on green and replaced backgrounds are both 7 ± 1 cm for a challenging 250-frame walking sequence with occlusions, and 7 ± 1 cm on a long sequence of 750 frames of gesturing and interaction. During gesturing, where arms are close to the camera, upper body (shoulder, elbow, wrist, finger) joint accuracy is higher than for the lower body (hip, knee, ankle, and toe) with 6 cm and 8 cm average error, respectively. During walking, upper and lower body error is similar with 7 cm. Please note that slight differences in skeleton topology between ground truth and EgoCap exist, which might bias the errors.

Despite the difficult viewing angle and image distortion of our egocentric setup, the overall 3D reconstruction error is comparable to state-of-the-art results of outside-in approaches [Rhodin et al. 2015, Elhayek et al. 2015, Amin et al. 2009, Sigal et al. 2010, Belagiannis et al. 2014], which reach 5–7 cm accuracy from two or more cameras, but only in small and open recording volumes, and for static cameras. In contrast, our algorithm scales to very narrow and cluttered scenes (see Figure 9) as well as to wide unconstrained performances (see Figure 8). No existing algorithm is directly applicable to these conditions and the strong distortions of the fisheye cameras, precluding a direct comparison. Closest to our approach is the fundamentally off-line inside-out method of Shiratori et al. [2011], who use 16 body-worn cameras facing outwards, reporting a mean joint position error of 2 cm on a slowly performed indoor walking sequence. Visually, their outdoor results show similar quality to our reconstructions, although we require fewer cameras, and can handle crowded scenes. It depends on the application whether head gear or body-worn cameras less impair the user's performance.

5.5 Model Components

Our objective energy consists of detection, color, smoothness, and pose prior terms. Disabling the smoothness term increases the reconstruction error on the validation sequences by 3 cm. Without the color term, accuracy is reduced by 0.5 cm. We demonstrate in the supplemental video that the influence of the color term is more significant in the outdoor sequences for motions that are very dissimilar to the training set. Disabling the detection term removes the ability to recover from tracking failures, which are usually unavoidable for fully automatic motion capture of long sequences with challenging motions. High-frequency noise is filtered with a Gaussian low-pass filter of window size 5.

6 Applications

We further evaluate our approach in three application scenarios with seven sequences of lengths of up to 1500 frames using *EgoRig1*, in addition to the three quantitative evaluation sequences. The captured users wear clothes not present in the training set. The qualitative results are best observed in the supplemental video.

6.1 Unconstrained/Large-Scale Motion Capture

We captured a `Basketball` sequence outdoors, which shows quick motions, large steps on a steep staircase, and close interaction of arms, legs and the basketball (supplemental video). We also recorded an outdoor `Walk` sequence with frequent arm-leg self-occlusions (Figure 8, right). With EgoCap, a user can even motion capture themselves while riding a bike in a larger volume of space (`Bike` sequence, Figure 8, left and center). The pedaling motion of the legs is nicely captured, despite frequent self-occlusions; the steering motion of the arms and the torso is also reconstructed. Even for very fast absolute motions, like this one on a bike, our egocentric rig with cameras attached to the body leads to little motion blur, which challenges outside-in optical systems. All this would have been difficult with alternative motion-capture approaches.

Note that our outdoor sequences also show the resilience of our method to different appearance and lighting conditions, as well as the generalization of our detector to a large range of scenes.

6.2 Constrained/Crowded Spaces

We also tested EgoCap with *EgoRig1* for motion capture on the `Crowded` sequence, where many spectators are interacting and occluding the tracked user from the outside (Figure 9). In such a

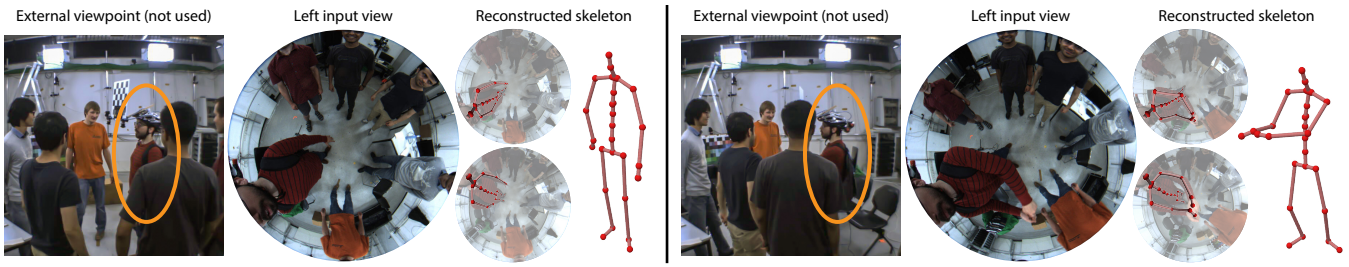


Figure 9: *Capturing social interaction in crowded scenes is of importance, but occlusions pose difficulties for existing outside-in approaches (left). The egocentric view enables 3D pose estimation, as demonstrated on the* `Crowded` *sequence. The visible checkerboard is not used.*

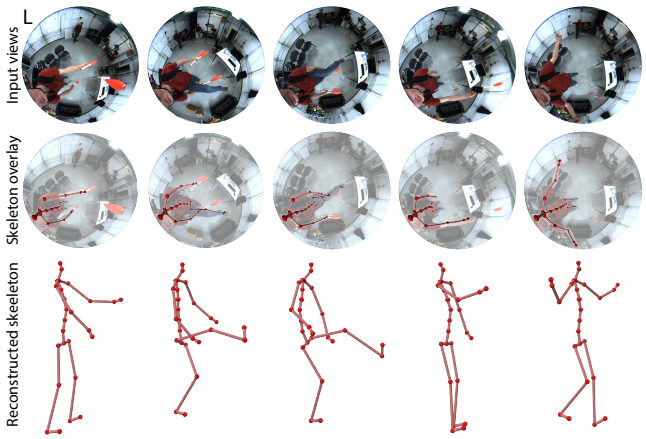


Figure 10: *Reconstruction results on the* `Juggler` *sequence, showing one input view and the estimated skeleton. Despite frequent self-occlusions, our approach robustly recovers the skeleton motion.*

setting, as well as in settings with many obstacles and narrow sections, outside-in motion capture, even with a dense camera system, would be difficult. In contrast, EgoCap captures the skeletal motion of the user in the center with only two head-mounted cameras.

The egocentric camera placement is well-suited for capturing human-object interactions too, such as the juggling performance `Juggler` (Figure 10). Fast throwing motions as well as occlusions are handled well. The central camera placement ensures that objects that are manipulated by the user are always in view.

6.3 Tracking for Immersive VR

We also performed an experiment to show how EgoCap could be used in immersive virtual reality (VR) applications. To this end, we use *EgoRig2* attached to an Oculus VR headset and track the motion of a user wearing it. We build a real-time demo application running at up to 15 fps, showing that real-time performance is feasible with additional improvements on currently unoptimized code. In this `Live` test, we only use color-based tracking of the upper body, without detections, as the detector code is not yet optimized for speed. The `Live` sequence shows that body motions are tracked well, and that with such an even more lightweight capture rig, geared for HMD-based VR, egocentric motion capture is feasible. In the supplemental video, we show an additional application sequence '`VR`', in which the the user can look down at their virtual self while sitting down on a virtual sofa. Current HMD-based systems only track the pose of the display; our approach adds motion capture of the wearer's full body, which enables a much higher level of immersion.

Global Pose Estimation For free roaming, the global rig pose can be tracked independently of external devices using structure-from-motion in the fisheye views. We demonstrate combined local and global pose estimation on the `Biking`, `Walk`, and `VR` sequence, using the structure-from-motion implementation of Moulon et al. [2013] provided in the OpenMVG library, see Figure 11 and the accompanying video. Such complete motion capture paves the way for immersive roaming in a fully virtual 3D environment.

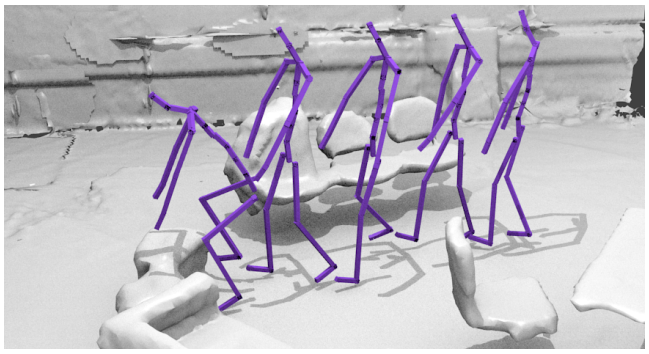


Figure 11: *Complete motion-capture example* `VR`*, in which our egocentric pose tracking is combined with global pose tracking using structure-from-motion, shown as a motion sequence in a 3D reconstruction of the scene. In a VR scenario, this would allow free roaming and interaction with virtual objects.*

7 Discussion and Limitations

We developed the first stereo egocentric motion-capture approach for indoor and outdoor scenes, that also works well for very crowded scenes. The combination of generative and detection-based pose estimation make it fare well even under poses with notable self-occlusions. Similar to other outside-in optical methods, tracking under occlusions by objects in the environment, such as a table, may lead to tracking failures. However, the detections enable our tracker to quickly recover from such occlusion failures. Interestingly, the egocentric fisheye camera setup provides stronger perspective cues for motion towards and away from the camera than with normal optics. The perspective effect of the same motion increases with proximity to the camera. For instance, bending an arm is a subtle motion when observed from an external camera, but when observed in proximity, the same absolute motion causes large relative motion, manifesting in large displacements and scaling of the object in motion.

The algorithm in this paper focuses on an entirely new way of capturing the full egocentric skeletal body pose, that is decoupled from global pose and rotation relative to the environment. Global pose can be inferred separately by structure-from-motion from the fisheye cameras or is provided by HMD tracking in VR applications. Fisheye cameras keep the whole body in view, but cause distortions

reducing the image resolution of distant body parts such as the legs. Therefore, tracking accuracy of the upper body is slightly higher than that of the lower body. Also, while overall tracking accuracy of our research prototype is still lower than with commercial outside-in methods, it shows a new path towards more unconstrained capture in the future. Currently, we have no real-time end-to-end prototype. We are confident that this would be feasible without algorithm redesign, yet felt that real-time performance is not essential to demonstrate the algorithm and its general feasibility.

Our current prototype systems may still be a bit bulky, but much stronger miniaturization becomes feasible in mass production; the design of *EgoRig2* shows this possibility. Some camera extension is required for lower-body tracking and might pose a problem with respect to social acceptance and ergonomics for some applications; However, we did not encounter practical issues during our recordings and VR tests, as users naturally keep the area in front of their head clear to not impair their vision. Moreover, handling changing illumination is still an open problem for motion capture in general and is not the focus of our work. For dynamic illumination, the color model would need to be extended. However, the CNN performs one-shot estimation and does not suffer from illumination changes. The training data also contains shadowing from the studio illumination, although extreme directional light might still cause inaccuracies. Additionally, loose clothing, such as a skirt, is not part of the training dataset and hence likely to reduce pose accuracy.

8 Conclusion

We presented EgoCap, the first approach for marker-less egocentric full-body motion capture with a head-mounted fisheye stereo rig. It is based on a pose optimization approach that jointly employs two components. The first is a new generative pose estimation approach based on a ray-casting image formation model enabling an analytically differentiable alignment energy and visibility model. The second component is a new ConvNet-based body-part detector for fisheye cameras that was trained on the first automatically annotated real-image training dataset of egocentric fisheye body poses. EgoCap's lightweight on-body capture strategy bears many advantages over other motion-capture methods. It enables motion capture of dense and crowded scenes, and reconstruction of large-scale activities that would not fit into the constrained recording volumes of outside-in motion-capture methods. It requires far less instrumentation than suit-based or exoskeleton-based approaches. EgoCap is particularly suited for HMD-based VR applications; two cameras attached to an HMD enable full-body pose reconstruction of your own virtual body to pave the way for immersive VR experiences and interactions.

Acknowledgements

We thank all reviewers for their valuable feedback, Dushyant Mehta, James Tompkin, and The Foundry for license support. This research was funded by the ERC Starting Grant project CapReal (335545).

References

AMIN, S., ANDRILUKA, M., ROHRBACH, M., AND SCHIELE, B. 2009. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*.

ANDRILUKA, M., PISHCHULIN, L., GEHLER, P., AND SCHIELE, B. 2014. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*.

BAAK, A., MÜLLER, M., BHARAJ, G., SEIDEL, H.-P., AND THEOBALT, C. 2011. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV*.

BELAGIANNIS, V., AMIN, S., ANDRILUKA, M., SCHIELE, B., NAVAB, N., AND ILIC, S. 2014. 3D pictorial structures for multiple human pose estimation. In *CVPR*.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*.

BREGLER, C., AND MALIK, J. 1998. Tracking people with twists and exponential maps. In *CVPR*.

BURENIUS, M., SULLIVAN, J., AND CARLSSON, S. 2013. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*.

CEREZO, E., PÉREZ, F., PUEYO, X., SERON, F. J., AND SILLION, F. X. 2005. A survey on participating media rendering techniques. *The Visual Computer 21*, 5, 303–328.

CHAI, J., AND HODGINS, J. K. 2005. Performance animation from low-dimensional control signals. *ACM Transactions on Graphics 24*, 3, 686–696.

CHEN, X., AND YUILLE, A. L. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*.

EGOCAP, 2016. EgoCap dataset. http://gvv.mpi-inf.mpg.de/projects/EgoCap/.

ELHAYEK, A., DE AGUIAR, E., JAIN, A., TOMPSON, J., PISHCHULIN, L., ANDRILUKA, M., BREGLER, C., SCHIELE, B., AND THEOBALT, C. 2015. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*.

FATHI, A., FARHADI, A., AND REHG, J. M. 2011. Understanding egocentric activities. In *ICCV*.

GALL, J., ROSENHAHN, B., BROX, T., AND SEIDEL, H.-P. 2010. Optimization and filtering for human motion capture. *International Journal of Computer Vision 87*, 1–2, 75–92.

HA, S., BAI, Y., AND LIU, C. K. 2011. Human motion reconstruction from force sensors. In *SCA*.

HE, K., ZHANG, X., REN, S., AND SUN, J. 2016. Deep residual learning for image recognition. In *CVPR*.

HOLTE, M. B., TRAN, C., TRIVEDI, M. M., AND MOESLUND, T. B. 2012. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing 6*, 5, 538–552.

INSAFUTDINOV, E., PISHCHULIN, L., ANDRES, B., ANDRILUKA, M., AND SCHIELE, B. 2016. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*.

JAIN, A., TOMPSON, J., ANDRILUKA, M., TAYLOR, G. W., AND BREGLER, C. 2014. Learning human pose estimation features with convolutional networks. In *ICLR*.

JAIN, A., TOMPSON, J., LECUN, Y., AND BREGLER, C. 2015. MoDeep: A deep learning framework using motion features for human pose estimation. In *ACCV*.

JIANG, H., AND GRAUMAN, K. 2016. Seeing invisible poses: Estimating 3D body pose from egocentric video. arXiv:1603.07763.

JOHNSON, S., AND EVERINGHAM, M. 2011. Learning effective human pose estimation from inaccurate annotation. In *CVPR*.

JONES, A., FYFFE, G., YU, X., MA, W.-C., BUSCH, J., ICHIKARI, R., BOLAS, M., AND DEBEVEC, P. 2011. Head-mounted photometric stereo for performance capture. In *CVMP*.

JOO, H., LIU, H., TAN, L., GUI, L., NABBE, B., MATTHEWS, I., KANADE, T., NOBUHARA, S., AND SHEIKH, Y. 2015. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*.

KIM, D., HILLIGES, O., IZADI, S., BUTLER, A. D., CHEN, J., OIKONOMIDIS, I., AND OLIVIER, P. 2012. Digits: Freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *UIST*.

KITANI, K. M., OKABE, T., SATO, Y., AND SUGIMOTO, A. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*.

LOPER, M., MAHMOOD, N., AND BLACK, M. J. 2014. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics 33*, 6, 220:1–13.

MA, M., FAN, H., AND KITANI, K. M. 2016. Going deeper into first-person activity recognition. In *CVPR*.

MEKA, A., ZOLLHÖFER, M., RICHARDT, C., AND THEOBALT, C. 2016. Live intrinsic video. *ACM Transactions on Graphics 35*, 4, 109:1–14.

MENACHE, A. 2010. *Understanding Motion Capture for Computer Animation*, 2nd ed. Morgan Kaufmann.

MOESLUND, T. B., HILTON, A., KRÜGER, V., AND SIGAL, L., Eds. 2011. *Visual Analysis of Humans: Looking at People*. Springer.

MOULON, P., MONASSE, P., AND MARLET, R. 2013. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*.

MURRAY, R. M., SASTRY, S. S., AND ZEXIANG, L. 1994. *A Mathematical Introduction to Robotic Manipulation*. CRC Press.

NEWELL, A., YANG, K., AND DENG, J. 2016. Stacked hourglass networks for human pose estimation. arXiv:1603.06937.

OHNISHI, K., KANEHIRA, A., KANEZAKI, A., AND HARADA, T. 2016. Recognizing activities of daily living with a wrist-mounted camera. In *CVPR*.

PARK, S. I., AND HODGINS, J. K. 2008. Data-driven modeling of skin and muscle deformation. *ACM Transactions on Graphics 27*, 3, 96:1–6.

PARK, H. S., JAIN, E., AND SHEIKH, Y. 2012. 3D social saliency from head-mounted cameras. In *NIPS*.

PFISTER, T., CHARLES, J., AND ZISSERMAN, A. 2015. Flowing ConvNets for human pose estimation in videos. In *ICCV*.

PISHCHULIN, L., INSAFUTDINOV, E., TANG, S., ANDRES, B., ANDRILUKA, M., GEHLER, P., AND SCHIELE, B. 2016. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*.

PONS-MOLL, G., BAAK, A., HELTEN, T., MÜLLER, M., SEIDEL, H.-P., AND ROSENHAHN, B. 2010. Multisensor-fusion for 3D full-body human motion capture. In *CVPR*.

PONS-MOLL, G., BAAK, A., GALL, J., LEAL-TAIXÉ, L., MÜLLER, M., SEIDEL, H.-P., AND ROSENHAHN, B. 2011. Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling. In *ICCV*.

PONS-MOLL, G., FLEET, D. J., AND ROSENHAHN, B. 2014. Posebits for monocular human pose estimation. In *CVPR*.

RHINEHART, N., AND KITANI, K. M. 2016. Learning action maps of large environments via first-person vision. In *CVPR*.

RHODIN, H., ROBERTINI, N., RICHARDT, C., SEIDEL, H.-P., AND THEOBALT, C. 2015. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*.

RHODIN, H., ROBERTINI, N., CASAS, D., RICHARDT, C., SEIDEL, H.-P., AND THEOBALT, C. 2016. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*.

ROGEZ, G., KHADEMI, M., SUPANCIC, III, J. S., MONTIEL, J. M. M., AND RAMANAN, D. 2014. 3D hand pose detection in egocentric RGB-D images. In *ECCV Workshops*.

SAPP, B., AND TASKAR, B. 2013. MODEC: Multimodal decomposable models for human pose estimation. In *CVPR*.

SCARAMUZZA, D., MARTINELLI, A., AND SIEGWART, R. 2006. A toolbox for easily calibrating omnidirectional cameras. In *IROS*.

SHIRATORI, T., PARK, H. S., SIGAL, L., SHEIKH, Y., AND HODGINS, J. K. 2011. Motion capture from body-mounted cameras. *ACM Transactions on Graphics 30*, 4, 31:1–10.

SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., AND BLAKE, A. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR*.

SIGAL, L., BĂLAN, A. O., AND BLACK, M. J. 2010. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87, 4–27.

SIGAL, L., ISARD, M., HAUSSECKER, H., AND BLACK, M. J. 2012. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision 98*, 1, 15–48.

SRIDHAR, S., MUELLER, F., OULASVIRTA, A., AND THEOBALT, C. 2015. Fast and robust hand tracking using detection-guided optimization. In *CVPR*.

STOLL, C., HASLER, N., GALL, J., SEIDEL, H.-P., AND THEOBALT, C. 2011. Fast articulated motion tracking using a sums of Gaussians body model. In *ICCV*.

SU, Y.-C., AND GRAUMAN, K. 2016. Detecting engagement in egocentric video. In *ECCV*.

SUGANO, Y., AND BULLING, A. 2015. Self-calibrating head-mounted eye trackers using egocentric visual saliency. In *UIST*.

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, 18:1–12.

TEKIN, B., ROZANTSEV, A., LEPETIT, V., AND FUA, P. 2016. Direct prediction of 3D body poses from motion compensated sequences. In *CVPR*.

THEOBALT, C., DE AGUIAR, E., STOLL, C., SEIDEL, H.-P., AND THRUN, S. 2010. Performance capture from multi-view video. In *Image and Geometry Processing for 3-D Cinematography*, R. Ronfard and G. Taubin, Eds. Springer, 127–149.

TOMPSON, J. J., JAIN, A., LECUN, Y., AND BREGLER, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*.

TOSHEV, A., AND SZEGEDY, C. 2014. DeepPose: Human pose estimation via deep neural networks. In *CVPR*.

URTASUN, R., FLEET, D. J., AND FUA, P. 2006. Temporal motion models for monocular and multiview 3D human body tracking. *Computer Vision and Image Understanding 104*, 2, 157–177.

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.

WANG, R. Y., AND POPOVIĆ, J. 2009. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics 28*, 3, 63.

WANG, J., CHENG, Y., AND FERIS, R. S. 2016. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*.

WEI, X., ZHANG, P., AND CHAI, J. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics 31*, 6, 188:1–12.

WEI, S.-E., RAMAKRISHNA, V., KANADE, T., AND SHEIKH, Y. 2016. Convolutional pose machines. In *CVPR*.

YANG, Y., AND RAMANAN, D. 2013. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 12, 2878–2890.

YASIN, H., IQBAL, U., KRÜGER, B., WEBER, A., AND GALL, J. 2016. A dual-source approach for 3D pose estimation from a single image. In *CVPR*.

YIN, K., AND PAI, D. K. 2003. Footsee: an interactive animation system. In *SCA*.

YONEMOTO, H., MURASAKI, K., OSAWA, T., SUDO, K., SHIMAMURA, J., AND TANIGUCHI, Y. 2015. Egocentric articulated pose tracking for action recognition. In *International Conference on Machine Vision Applications (MVA)*.

ZHANG, P., SIU, K., ZHANG, J., LIU, C. K., AND CHAI, J. 2014. Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM Transactions on Graphics 33*, 6, 221:1–14.

A Implementation Details

Color Dissimilarity For measuring the dissimilarity $d(\mathbf{m}, \mathbf{i})$ of model color $\mathbf{m}$ and image pixel color $\mathbf{i}$ in Equation 5, we use the HSV color space (with all dimensions normalized to unit range) and combine three dissimilarity components:

1. For saturated colors, the color dissimilarity d_s is computed using the squared (minimum angular) hue distance. Using the hue channel alone gains invariance to illumination changes.

2. For dark colors, the color dissimilarity d_d is computed as twice the squared value difference, i.e. $d_d(\mathbf{m}, \mathbf{i}) = 2(m_v - i_v)^2$. Hue and saturation are ignored as they are unreliable for dark colors.

3. For gray colors, the distance d_g is computed as the sum of absolute value and saturation difference, i.e. $d_g(\mathbf{m}, \mathbf{i}) = |m_v - i_v| + |m_s - i_s|$. Hue is unreliable and thus ignored.

We weight these three dissimilarity components by $w_s = \sqrt{m_s}/Z$, $w_d = \max(0, 0.5 - m_v)/Z$ and $w_g = \max(0, 0.5 - m_s)/Z$ respectively, where Z normalizes the sum of these weights to unity. The total dissimilarity is computed by $d(\mathbf{m}, \mathbf{i}) = \phi(w_s d_s + w_d d_d + w_g w_g)$ where $\phi(x) = 1 - (1 - x)^4(8x + 2)$ is a smooth step function. We employ a two-sided energy, i.e. E_{color} can be negative: For dissimilar colors, $d \approx 1$ and approaches -1 for similar colors.