Full-Body Human Motion Capture from Monocular Depth Images

Thomas Helten[1], Andreas Baak[1], Meinard Müller[2], and Christian Theobalt[1]

[1] MPI Informatik, Campus E1.4, 66123 Saarbrücken, Germany
{thelten,abaak,theobalt}@mpi-inf.mpg.de
[2] International Audio Laboratories, Am Wolfsmantel 33, 91058 Erlangen, Germany
meinard.mueller@audiolabs-erlangen.de

Abstract. Optical capturing of human body motion has many practical applications, ranging from motion analysis in sports and medicine, over ergonomy research, up to computer animation in game and movie production. Unfortunately, many existing approaches require expensive multi-camera systems and controlled studios for recording, and expect the person to wear special marker suits. Furthermore, marker-less approaches demand dense camera arrays and indoor recording. These requirements and the high acquisition cost of the equipment makes it applicable only to a small number of people. This has changed in recent years, when the availability of inexpensive depth sensors, such as time-of-flight cameras or the Microsoft Kinect has spawned new research on tracking human motions from monocular depth images. These approaches have the potential to make motion capture accessible to much larger user groups. However, despite significant progress over the last years, there are still unsolved challenges that limit applicability of depth-based monocular full body motion capture. Algorithms are challenged by very noisy sensor data, (self) occlusions, or other ambiguities implied by the limited information that a depth sensor can extract of the scene. In this article, we give an overview on the state-of-the-art in full body human motion capture using depth cameras. Especially, we elaborate on the challenges current algorithms face and discuss possible solutions. Furthermore, we investigate how the integration of additional sensor modalities may help to resolve some of the ambiguities and improve tracking results.

1 Introduction

The recording and analysis of full-body human motion data constitutes an important strand of research in computer vision, computer graphics and many related fields of visual computing. Full body human motion capture has many applications in divers areas, ranging from character animation for movie and game productions, sports sciences, and human computer interaction. Unfortunately, the methods for measuring human skeletal motion that were available until recently impose stark constraints on applicability and can lead to high acquisition cost. Most applications in the movie and game industry, medical research and rehabilitation, as well as sports sciences are often based on optical marker-based

M. Grzegorzek et al. (Eds.): Time-of-Flight and Depth Imaging, LNCS 8200, pp. 188–206, 2013.

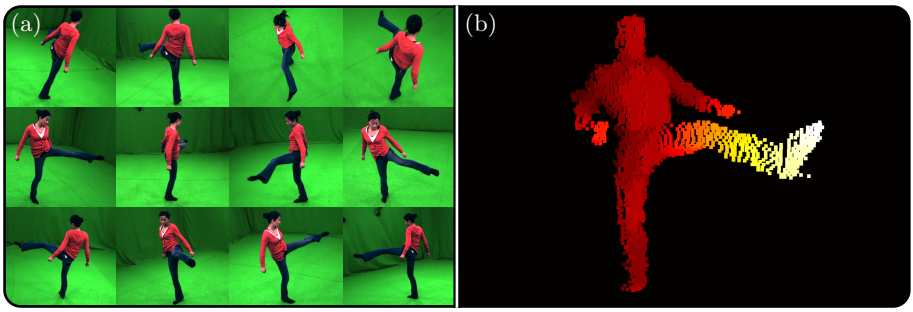

Fig. 1. (a) Input color images for a typical markerless multi-camera motion capture approach. **(b)** Input depth image for a typical depth tracking approach.

or marker-less approaches, see [1] for an overview. These approaches often need multi-view input images, recorded in controlled environments using expensive and calibrated recording equipment, see also Fig. 1a. These requirements render them unaffordable for many users, or even completely unsuitable, such as in home user applications.

In the recent years, depth sensing devices such as time-of-flight (ToF) cameras or the Microsoft Kinect have triggered a new strand of research, where human motion data is inferred from so called 2.5D depth maps. Such cameras are easy to set-up and are inexpensive compared to the systems required by the approaches above. The provided data is especially appealing for tracking because of two reasons. Firstly, it is more resilient to challenging surface and appearance properties of objects and in most cases independent from controlled lighting conditions. Secondly, the provided depth maps enable easier background subtraction and provide rich geometric information even when using only a single camera, see also Fig. 1b. In consequence, several algorithms were introduced recently that can capture full body human skeletal poses from a single depth camera view. While they do not yet reach the same level of accuracy as classical multi-camera-based approaches, many of them perform in real-time and have paved the trail for some new interaction applications in home user environments.

However, despite the advances in this field, there are still many fundamental algorithmic obstacles to overcome in order to bridge the immense quality and robustness gap between depth-camera based tracking and earlier multi-camera approaches. Current algorithms are challenged by the non-trivial noise characteristics of depth cameras. Understanding and characterizing this noise (see also chapter "Denoising Strategies for Time-of-Flight Data") and properly accommodating for it (see also chapter "Stabilization of 3D Position Measurement") in the pose estimation methods is thus a key requirement. Another set of challenges originates from the fact that depth images are very sparse. While already with multiple available camera views the process of inferring pose from images is highly ambiguous, this problem is even more difficult in monocular pose reconstruction. Algorithms are challenged by occlusions resulting in missing information. Another example is the fact that the orientation of rotationally symmetric body parts, such as arms and legs, is ambiguous in the depth data.

In this article, we want to give an overview on the current state-of-the art in human pose estimation from depth images, see Sect. 2. We will review the advantages and disadvantages of the main categories of algorithmic strategies for monocular pose estimation from depth, which includes generative and discriminative strategies. We will also put a focus on the basic principle of so-called hybrid trackers that combine these two tracking recipes. Based on this review of state-of-the-art, we will elaborate on primary algorithmic limitations and challenges that current methods have to overcome, and present ideas and an outlook to possible ways of achieving this, see Sect. 3. In particular, we will use the example approach presented by Baak *et al.* [2] as instructional example, see Sect. 3.4.

2 State-of-the-Art

Nowadays, most commercial solutions to full-body human motion capture employ techniques that are invasive to the scene. Some approaches are based on mechanical or electronic exoskeletons, or other external sensors placed on the body. But the most widely used techniques require the person to wear special suits with retro-reflective markers whose motion is picked up by a multi-camera system to compute the skeletal motion of the person [3]. Due to the complex apparatus, these approaches are expensive, need a lot of preparation time, and are restricted to controlled recording environments which constrains their application to specialized professional users. To overcome this limitation, researchers in computer vision and computer graphics started to develop marker-less skeletal pose estimation algorithms. They can capture skeletal motion from multi-view video of a moving person, without needing markers in the scene. An extensive overview of these methods is beyond the scope of this chapter, and a review can be found in [4], but the main concepts are as follows. Most approaches use some form of 3D kinematic skeleton model augmented by shape primitives, such as cylinders [5], a surface mesh [6,7,8], or probabilistic density representations attached to the human body [9]. Optimal skeletal pose parameters are often found by minimizing an error metric that assesses the similarity of the projected model to the multi-view image data using features. Local optimization approaches are widely used due to their high efficiency, but they are challenged by the highly multimodal nature of the model-to-image similarity function [9,8]. Global pose optimization methods can overcome some of these limitations, however at the price of needing much longer computation times [10,6]. Some approaches aim to combine the efficiency of local methods with the reliability of global methods by adaptively switching between them [6]. Even though marker-less approaches succeed with a slightly simpler setup, many limitations remain: computation time often precludes real-time processing, recording is still limited to controlled settings, and people are still expected to wear relatively tight clothing. Furthermore, marker-less motion capture methods deliver merely skeletal motion parameters.

In contrast, marker-less performance capture methods go one step further and reconstruct deforming surface geometry from multi-view video in addition to skeletal motion. Some methods estimate the dynamic scene geometry using variants of shape-from-silhouette methods or combinations of shape-from-silhouette

and stereo[11,12,13,14], but in such approaches establishing space-time coherence is difficult. Template-based methods deform a shape template to match the deformable surface in the real scene, which implicitly establishes temporal coherence [15,16], also in scenes with ten persons. All the developments explained so far aim towards the goal of high-quality reconstruction, even if that necessitates complex and controlled indoor setup. In contrast, depth-based tracking of full-body human motion focuses on using inexpensive recording equipment that is easy to setup and to use in home user applications. As a consequence, depth based have to deal with various challenges that marker-less tracking approaches do not face. Commercial systems that make use of this kind of motion tracking can be found *e. g.* in the Microsoft Kinect for XBox[1], the SoftKinetic IISU Middleware[2] for pose and gesture recognition, as well as the SilverFit[3] system for rehabilitation support. So far, several depth-based tracking methods have been published that can be classified into three basic types: Generative approaches, discriminative approaches and hybrid approaches. In this chapter, we give a general overview over full-body tracking approaches. We refer to the chapter "A Survey on Human Motion Analysis from Depth Data" for activity recognition and body part motion in general. Furthermore, we refer to the chapter "Gesture Interfaces with Depth Sensors" for the specific case of hand and arm motion tracking. The later chapter also discusses a special kind of generative tracking approach which makes use of so-called self-organizing maps (SOM).

2.1 Generative Approaches

Generative approaches use parametrized body models that are fit into the depth data using optimization schemes. In particular, the optimization process maximizes a model-to-image consistency measure. This measure is hard to optimize due to the inherent ambiguity in the model-to-data projection. In particular, when using monocular video cameras, this ambiguity precludes efficient and reliable inference of a usable range of 3D body poses. Depth data reduce this ambiguity problem but it is still one of the main algorithmic challenges to make generative methods succeed.

A first approach for obtaining pose and surface of articulated rigid objects from ToF depth images was presented in [17]. Under the assumption that the movement of the tracked object is small *w. r. t.* the capture speed of the depth camera, the authors track individual bones from a manually pre-labeled depth image using an iterative closest point (ICP) approach. In each frame, previously unlabeled depth pixels are assigned to the bone that best explains the unlabeled depth pixel. However, this approach was not real-time capable, running at around 0.5 frames per second (FPS). Another approach [18] that is specialized on human motion, generates point correspondences for an ICP based optimization from both 3D and 2D input. An example for 2D input could be a body part

[1] `http://www.xbox.com/Kinect`
[2] `http://www.softkinetic.com`
[3] `http://www.silverfit.nl/en.html`

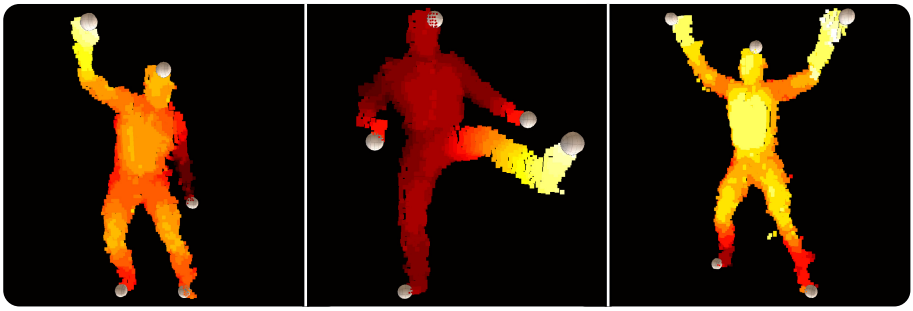

Fig. 2. First five geodesic extrema (white spheres) computed for several poses. These five extrema typically correspond to the four end-effectors (two hands, two feet) and the head of the person.

or feature detector working on 2D color images. All 3D points that could be projected onto the 2D feature point now define a ray in 3D space. The closest point of this ray to the model is used to generate a traditional 3D point constraint. The authors report a performance of 25 fps with this method, but the approach is limited to simple non-occluded poses since otherwise the tracker would converge to an erroneous pose minimum from which it cannot recover. Another early approach for real time capable depth-based motion tracking from monocular views was presented in [19]. Here, the authors describe a general pipeline for obtaining pose parameters of humans from a stream of depth images that are then used to drive the motion of a virtual character in *e. g.* video games. To further increase the performance of generative approaches [20] proposed porting the computational intense local optimization to the graphics processor. However, all these approaches tend to fail irrecoverably when the optimization is stuck in a local minimum. This problem also exists in other vision-based approaches and was *e. g.* discussed in [21]. In general, these tracking errors occur due to the ambiguous model-to-data mapping in many poses, as well as fast scene motion. While the latter problem can be remedied by increasing the frame rate, the former was addressed by more elaborated formulations of the energy function. One option was lately presented in [22], where the authors proposed a modified energy function that incorporates empty space information, as well as inter-penetration constraints. A completely different approach was shown in [23]. Here, multiple depth cameras were used for pose estimation which reduces the occlusion problem and enabled capturing the motion of multiple person using high resolution body models. The approach is not real-time capable, though. With all these depth-based methods, real-time pose estimation is still a challenge, tracking may drift, and with exception to [23], the employed shape models are rather coarse which impairs pose estimation accuracy.

2.2 Discriminative Approaches

On the other hand, discriminative approaches focus on detecting certain features in the depth data—such as joint locations—and later combine these independent

cues to form a body pose hypothesis. These feature are often learned for a predefined set of poses. For this reason, discriminative methods are not dependent on a numerical optimization procedure, and can infer pose also without temporal context and continuity. One algorithm for detecting human body parts in depth images was presented in [24]. Here, the authors use so-called geodesic extrema calculated by iteratively using Dijkstra's algorithm on a graph deduced by connecting all depth pixels in the 2.5D depth data into a map. The assumption here is that geodesic extrema generally align with salient points of the human body, such as the head, the hands, or the feet, see also Fig. 2. To label the retrieved geodesic extrema according to the corresponding body part, the authors employ local shape descriptors on normalized depth image patches centered at the geodesic extrema's positions. Another body part detection approached is pursued in [25], where the authors deduce landmark positions from the depth image and include regularizing information from previous frames. These positions are then used in a kinematic self retargeting framework to estimate the pose parameters of the person. In contrast, the approach described in [26] uses regression forest learned on simple pair-wise depth features to do a pixel-wise classification of the input depth image into body parts. To obtain a working regression forest for joint classification that works under a large range of poses, though, the authors had to train the classifier on approx. 500 000 synthetically generated and labeled depth images. For each body part, joint positions are then inferred by applying a mean shift-based mode finding approach on the pixels assigned to that body part, see also Fig. 3a. Using also regression forests for body part detection, [27] determine the joint positions by letting each depth pixel vote for the joint positions of several joints. After excluding votes from too distant depth pixels and applying a density estimator on the remaining votes, even the probable positions of non-visible joints can be estimated, see also Fig. 3b. Finally, [28] generate correspondences between body parts and a pose and size parametrized human model, which they also achieve by using depth features and regression forests. The parameters of this model are then found using a one shot optimization scheme, *i. e.* without iteratively recomputing the established correspondences. Discriminative approaches show impressive tracking results, where some discriminative methods even succeed in detecting joint information also in non-frontal occluded poses. However, since they often detect features in every depth frame independently, discriminative approaches tend to yield temporally unstable pose estimations results. Furthermore, for many learning-based methods, the effort to train classifiers can be significant.

2.3 Hybrid Approaches

Combining the ideas of generative and discriminative approaches, hybrid approaches try to harness the advantages from both tracker types. On the one hand, hybrid trackers inherit the stability and temporal coherence of pose estimation results common to generative trackers. On the other hand, they show the robustness of pose inference even in partly occluded poses that characterizes discriminative approaches. A first method, in the domain of 3D surface

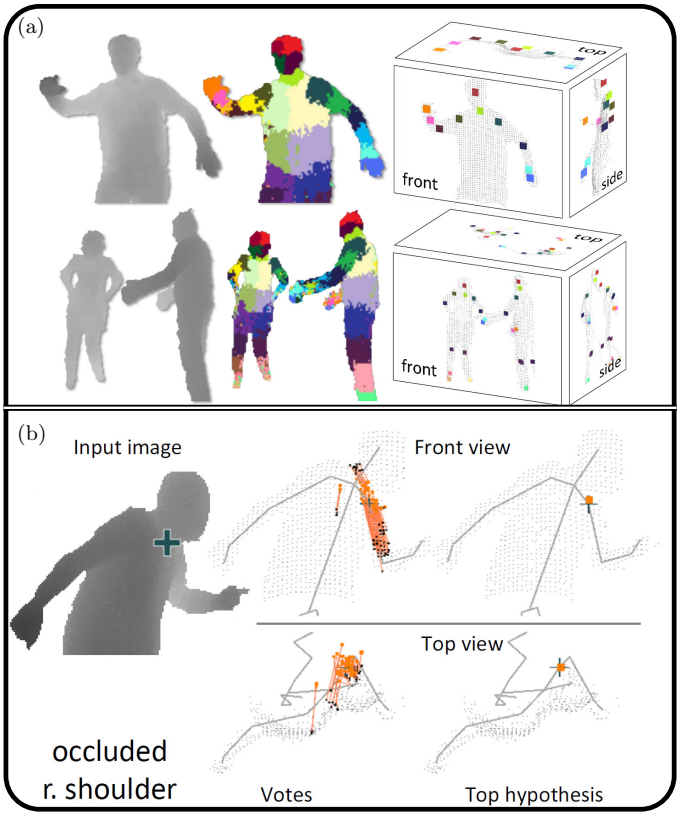

Fig. 3. Regression-forest-based discriminative trackers. The images were taken from the respective papers. **(a)** Body part and joint detection as presented in [26]. **(b)** Voting approach for occluded joints as described in [27].

reconstruction, was presented in [29]. Here, the discriminative tracker is used for initializing the surface model, while the generative tracker enforces the observance of distance constraints. The authors also sketched, how their approach can be applied to human pose reconstruction. At the same time, the first method with specialization to human pose estimation was presented in [30]. This work combines the geodesic extrema-based body part recognition presented in [24] with a generative pose optimization scheme based on articulated ICP. Furthermore, the authors introduce a dataset comprising of calibrated ToF depth images and ground-truth marker positions that serves as common benchmark for future work in that field. The works by Baak *et al.* [2] and Ye *et al.* [31] also use a discriminative tracker to initialize a generative pose estimation algorithm. In detail, the approach presented in [31] uses a database consisting of 19 300 poses. For each of these poses, four synthesized depth images were rendered from different views. Using a principal axis based normalization, the point clouds are indexed using their coefficients in a PCA subspace. Here, the normalization of

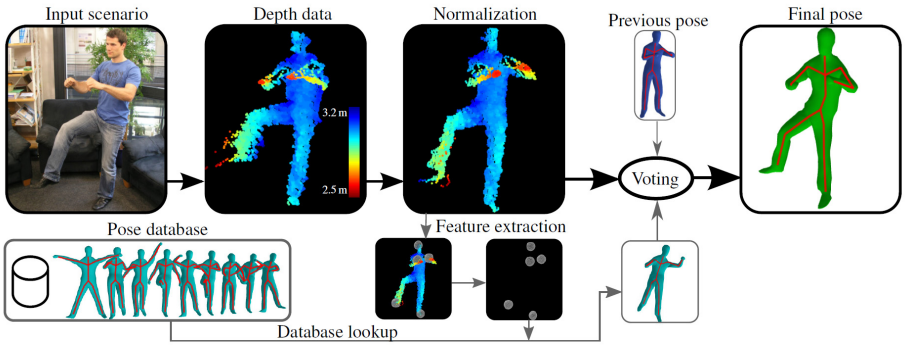

Fig. 4. Schematic overview of a hybrid depth tracker as suggested by Baak et al. [2]

the point cloud in combination with the rendering from four different views is used to retrieve poses from the database independent from the orientation *w. r. t.* the depth camera. Note that by storing four different views in the database, the index size is increased to 77 200, while still only 19 300 poses are contained in the database. During tracking, the input point cloud is normalized in the same way, its PCA-coefficients are calculated and used for retrieving a similar point cloud in the database. Finally, they refine the retrieved pose using the Coherent Drift Point algorithm presented in [32]. This approach shows good pose estimation results on the benchmark dataset introduced in [30]. However, their approach does not run in real time—inferring the pose in one frame takes between 60 s and 150 s.

In contrast, the approach showcased in [2] uses a modified iterated version of Dijkstra's algorithm to calculate geodesic extrema similar to the approach in [24]. The stacked positions of the first five geodesic extrema, which often co-align with the head, hands and feet, serve as index into a pose database consisting of 50 000 poses. The suitability of such an approach has been previously discussed in [33], where the authors used the stacked positions of the body's extremities (head, hands, and feet) to index a database containing high dimensional motion data. As index structure the authors employed a kd-tree facilitating fast nearest neighbor searches. To be invariant to certain orientation variations of the person, Baak *et al.* normalize the query and the database poses based on information deduced from the depth point cloud. The incorporated generative tracker is a standard ICP approach that builds correspondences between preselected points from the parametrized human model and points in the depth point cloud. In each frame, they conduct two local optimizations, one initialized using the pose from the previous frame and one using the retrieved pose from the pose database. Using a late fusion step they decide based on a sparse Hausdorff-like distance function which pose obtained from the two local optimizations best describes the observed depth image. This pose is then used as final pose hypothesis, see Fig. 4 for an overview of their approach. While not showing as good results as the approach presented in [31], their tracker runs much faster at around $50-60$ frames

per second, enabling very responsive tracking. Another real-time approach was recently proposed by *e. g.* [34]. Here, the authors use a discriminative body-part detector similar to [26] to augment a generative tracker. In particular, they use the pose obtained from the discriminative tracker only for initialization at the beginning of the tracking and for reinitializing the generative tracker in cases of tracking errors. For detecting wrongly tracked frames, they measure how well their body model with the current pose parameters explains the observed point cloud. Hybrid approaches, harnessing the advantages of both tracking worlds, are able to show superior performance compared too purely discriminative or generative approaches. However, even the current state-of-the-art hybrid trackers still have limitations, which we will elaborate on in the following.

3 Open Challenges and Possible Solutions

While providing good overall tracking results, hybrid approaches still suffer from the noisy character and the sparsity of the depth data and are prone to ambiguities originating from occlusions. In this section, we will discuss the various challenges current approaches still face, elaborate on the reasons, and give an outlook how these problems could be approached. For the special case of denoising depth data we refer to the chapter "Denoising Strategies for Time-of-Flight Data".

3.1 Accuracy of the Body Model

Most trackers use an underlying model of the human body. Such models vary drastically ranging from simple representations as graphs [17,25,26,27,28,29,31], over articulated rigid bodies [18,20,22,34] to complex triangle meshes driven by underlying skeletons using skinning approaches [2,23,30]. Here, the complexity of the model mainly depends on the intended application. While some approaches are only interested in tracking specific feature points of the body such as the positions of the extremities [24] or joint positions [26], other approaches try to capture pose parameters such as joint angles [2,22,28,30,31,34], or even the complete surface of the person including cloth wrinkles and folds [23]. Another requirement for a detailed surface model may be the energy function used in generative or hybrid approaches. In particular, ICP-based trackers benefit from an accurate surface model to build meaningful correspondences between the model and the point cloud during optimization. In order to circumvent the problem of obtaining an accurate model of each individual person, some approaches use a fixed body model and scale the input data instead [2]. However, this approach fails for persons with very different body proportions.

In general, the model of the tracked person is often assumed to be created in a pre-processing step using manual modeling or special equipment as full-body laser scanners. While this is a viable way in movie and game productions or in most scientific settings, in home user scenarios it is not feasible. To this end, most algorithms applied in home user scenarios, such as [26] use a different

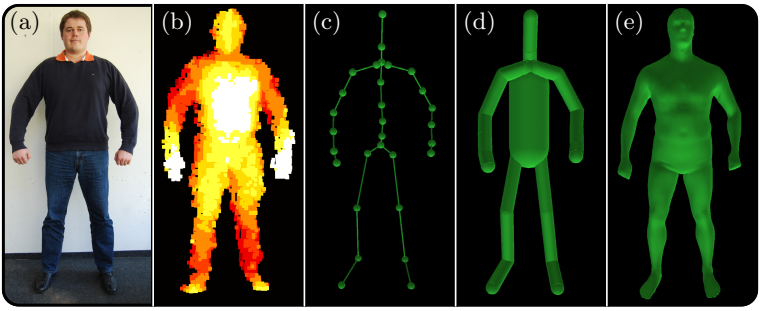

Fig. 5. **(a)** Body shape of a person to be tracked. **(b)** Depth image of shape. **(c)** Graph model. **(d)** Model based on articulated cylinders and spheres . **(e)** High resolution surface model.

approach. In a pre-processing step the authors use a large number of body models of different sizes and proportions to learn a decision-forest-based classifier that is able to label depth pixels according to the body part they belong to. As a consequence, this classifier becomes invariant to the size of the person and its proportions. During the actual tracking, the learned classifier can be used without obtaining an actual body model of the tracked person. Based on the labeled depth pixel the authors employ a heuristic to deduce the most probable joint position. This approach runs in real-time and works for many tracking applications.

However, for some augmented reality applications the reconstruction quality obtained from simple graphical body models may not be sufficient enough. A popular example is virtual try-on, where the person can wear a piece of virtual apparel that plausibly interacts with the person's body motion. Here, an accurate reconstruction of the person's body surface is beneficial in order to ensure believable visual quality or to give good indication whether the cloth actually fits. One possible approach would be to infer a high resolution body model from depth data in a pre-processing step and then use this model for tracking, visualization or physical simulations of objects in the augmented scene. Recently, one approach [35] has addressed this issue. Here, the authors fit a pose and shape parametrized model into the depth point clouds using an ICP-based approach. The point clouds were obtained from four sequentially captured depth images showing the person from the front, the back and two sides. However, the fact that the person had to reproduce the same pose in all four images and the optimization's runtime of about one hour makes this approach not applicable in home user scenarios. For an explanation how to obtain a pose and shape parametrized model, we refer to [36,37].

3.2 Rotational Ambiguities

Another inherent challenge to all depth-based trackers are rotational ambiguities. Depth data contains rich information about the relative location of objects

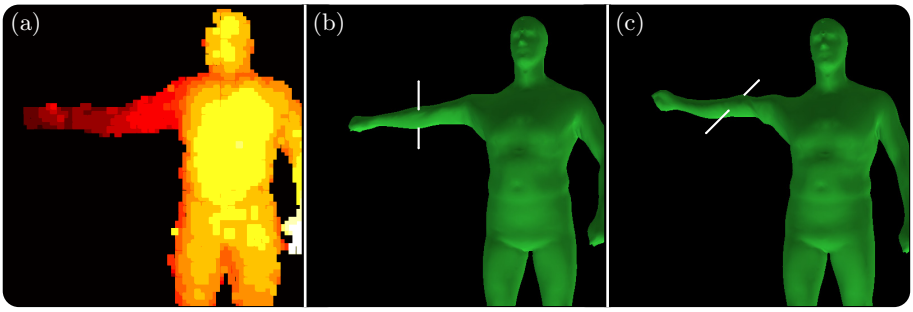

Fig. 6. Rotational ambiguities of depth data. **(a)** Input depth image. **(b)** One typical output from a generative pose estimation procedure. Note that the axis of the elbow joint is vertical. **(c)** Another possible output, the axis of the elbow joint is now horizontal.

which enables easy background subtraction compared to vision based approaches on intensity images. However, depth images reveal only little information about the surface structure and no color information at all. This makes it hard to determine the correct orientation of rotational symmetric objects, such as the body extremities. Since most depth trackers only depend on very simplistic underlying body models with isotropic extremities [18,20,22,34] or even graphs [17,25,26,27,28,29,31] that do not have any volume at all, they can simply ignore the aforementioned problem. However, these trackers also do not provide any pose information about the twist of the arms or the legs. In contrast, trackers that use complex triangle meshes for defining the body's surface [2,23,30] should not ignore rotational ambiguities. In particular, for these approaches the used generative tracker might come to different results depending on its initialization. An example can be seen in Fig. 6. Here, the depth image shown in Fig. 6a reveals only little information on how the arm is oriented. Two possible solutions of a generative tracker are depicted in in Fig. 6b&c. The difference between both solutions lies in the twist of the arm. While in Fig. 6b the axis of the right elbow joint is oriented vertically, it is oriented horizontally in Fig. 6c. In this example, the latter would semantically be the correct pose estimation result. At first glance this might not have huge impact on the overall performance of the tracker. However, a tracking error might serve as initialization for the next frame. Lets consider the scenario that the tracked person bends her arm with the forearm pointing upwards. While this is a straight-forward task for the generative tracker initialized with the pose shown in Fig. 6c, a local optimization starting with the pose shown in Fig. 6b is more likely to get stuck in a local minimum. Unfortunately, none of the presented trackers employs methods to prevent this. While pure generative trackers are likely to fail in such situations and may not be able to proceed, discriminative trackers completely avoid this issue by tracking each frame independently and not relying on local optimization. In contrast, hybrid approaches, such as presented in [2,34], detect the failure of their generative tracker and reinitialize it using pose estimations of their discriminative tracker.

Similar challenges are also faced in other tracking fields as *e. g.* marker-less motion capture. Here, so called silhouetted-based trackers that estimate the pose of the person from multiple, binary (foreground vs. background) images, suffer from the same challenge being unable to determine the correct orientation of the person's extremities. One approach to tackle this was presented in [7], where the authors included information from another sensor modality to correctly detect the orientation of the extremities independent from ambiguous optical information. In particular, their approach relies on orientation data obtained from five inertial sensors attached to the lower legs, forearms and the trunk of the person. By including the measured orientations into the energy function of their generative approach, tracking errors in rotationally symmetric limbs could be avoided.

3.3 Occlusions

The third and by far greatest challenge for today's depth trackers are occlusions. Occlusions stem from the fundamental principle how depth images (and other optical data) is obtained. Light is reflected by some object and detected by some light sensitive sensor inside the camera. If light from an object, *e. g.* a body part, cannot reach the sensor of the camera because another object in between, the object is occluded. As a consequence, one cannot obtain any usable information about the occluded object. Present depth trackers deal with occlusions in various ways. Some trackers simply avoid this by requiring the tracked person to strike only poses where all body parts are clearly visible to the depth camera [2,30,34]. Such trackers often show undefined behavior if the requirements are not met, see Fig. 7 for some representative failure cases. Some discriminative trackers allow for non frontal poses but do not give any pose hypothesis for non-visible parts [25,26,28,34]. In contrast, the approach presented in [27] uses a regression forest-based approach to learn the relative joint positions for a depth pixel based on depth values in its neighborhood. Calculating the density mean on a set of votes yields a hypothesis even for occluded joints. As most learning based approaches, this approach shows good results on poses close to the one used for learning and vice versa. In a pure generative setting, the approach proposed in [22] includes two additional constraints into the energy function to produce plausible results for occluded body parts. The first constraint prevents body parts from entering empty space, *i. e.* parts in the depth image where no foreground pixels were detected. The second constraint prevents body parts from inter-penetrating. However, without an actual measurement it is impossible to deduce the correct pose for occluded body parts.

We see two ways that could help tracking in difficult scenes. Firstly, occlusions could be reduced by dynamically moving the cameras during the recording of the scene. Secondly, occlusions could be handled by adding another input modality that does not depend on visual cues. As for the first approach, the authors in [23] make use of three Kinect depth cameras that are carried by operators around a scene. At a given frame, the depth input of the three Kinects is then fused into one point cloud representation of the whole scene. Using a generative

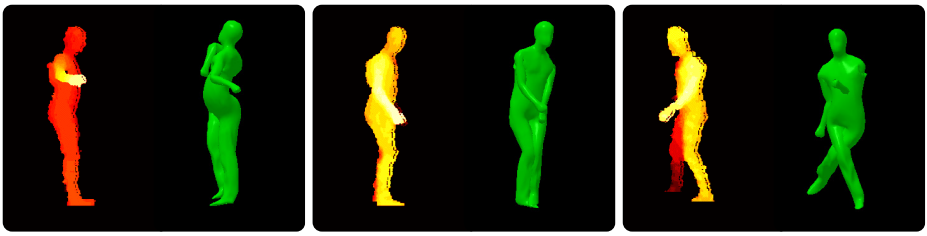

Fig. 7. Illustration of typical tracking artifacts in the case of non-frontal poses and occlusions. Many trackers require the tracked person to face the depth camera and have all arms and legs clearly visible. If those requirements are not met, this results in strong tracking artifacts. These example images where generated using the approach presented in [2].

tracking approach, the poses of the persons are tracked by fitting a rigged surface mesh into the point cloud. While this approach shows good results even for multiple persons in close contact, the runtime of the approach is not real-time and the use of multiple Kinect cameras is not feasible in home user scenarios. Furthermore, the use of several Kinect cameras simultaneously bears its own challenge since these cameras, in contrast to color cameras, interfere with each other's measurement. In order to reduce the interference of multiple Kinects, the authors of [38,39] applied vibration patterns to each camera. These vibrations have the effect that the point pattern projected by one Kinect looks blurred when seen from a different Kinect. In contrast, the pattern does not look blurred for the Kinect it is projected from, since its projector is moved in the same way its camera is. A similar effect is achieved in the approach presented in [23], since the three Kinects are not installed on tripods but hand-held by the camera operators. However, even when using multiple depth cameras, occlusions are difficult to prevent in many tracking scenarios.

As for the second approach, the fusion of different sensor modalities has become a successful approach for dealing with challenging tasks, in other research fields. An approach combining two complementary sensor types for full body human tracking in large areas was presented in [40]. Here, densely placed inertial sensors, one placed on every limb of the body, provide an occlusion independent estimation of the persons body configuration using measured global orientations. Since inertial sensors cannot measure their position, this information is provided by an optical system mounted to a robot accompanying the tracked person. Unfortunately, their approach does not include the rich optical information for supporting the tracking of the persons body configuration. Their approach rather solves two independent sub task, determining the local body configuration and estimating the global position of the person.

At this point, we want to take a second look on the approach presented in [7], which we also discussed in Sect. 3.2. In this approach, the main intention of using inertial sensors in a classical marker-less tracking framework was to prevent erroneous tracking that stems from the ambiguous representation of body

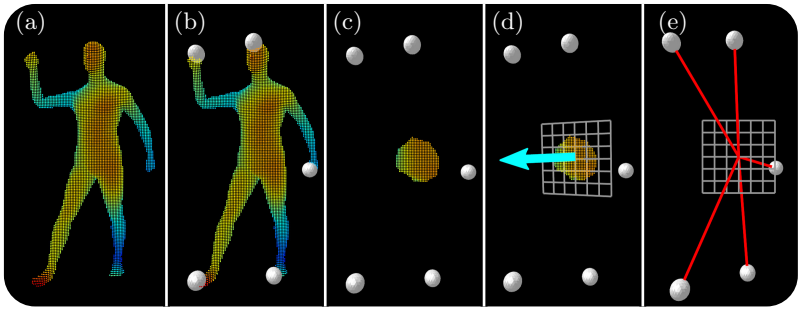

Fig. 8. Normalization of the query pose as presented in [2]. **(a)** Input point cloud of the tracked person. **(b)** Detected end effector positions. **(c)** Segmentation of the torso using mean-shift approach. **(d)** Plane fitted into torso points. The normal of the plane determines the front direction. **(e)** Normalized (front direction pointing towards camera) end effector positions as used for querying.

extremities in silhouette images. Another interesting side-effect is that the inertial sensors provide information about the limb orientations even in situations when the limbs are not visible to the camera. While in the presented scenario this effect was not important because multiple cameras enabled an almost occlusion free observation of the tracked person, this effect might be very important in monocular tracking approaches. In particular, many current depth-based trackers would benefit from additional information that does not depend on visual cues. In the following, we will take a state-of-the-art depth tracker and explain in detail how inertial information could be included to increase the performance in challenging tracking situations.

3.4 Improvement of a Hybrid Tracking Approach

The hybrid depth tracker presented by Baak *et al.* [2] states a typical example for combining a generative (local optimization) approach with a discriminative (DB lookup) approach. While their real-time tracking approach shows good performance on fast and dynamic motions, the tracker requires the person to face the camera during tracking. Furthermore, if body parts are occluded, the tracker might produce erroneous tracking results, see also Fig. 7. In this section, we elaborate on some of the limitations of this approach and discuss modifications to enhance its tracking performance. Furthermore, we will show that including additional complementary sensor information, such as provided by inertial sensors, may support the tracking in challenging tracking situations.

The requirement for frontal poses stems from design decisions made by the authors. In particular, the authors employ a database with normalized poses that serve as initialization to the generative tracker. As query to the database, the authors employ so called geodesic extrema, inspired by [24], computed on the depth point cloud that often co-align with salient features of the persons body such as the head, hands and feet. The normalization of the database was chosen

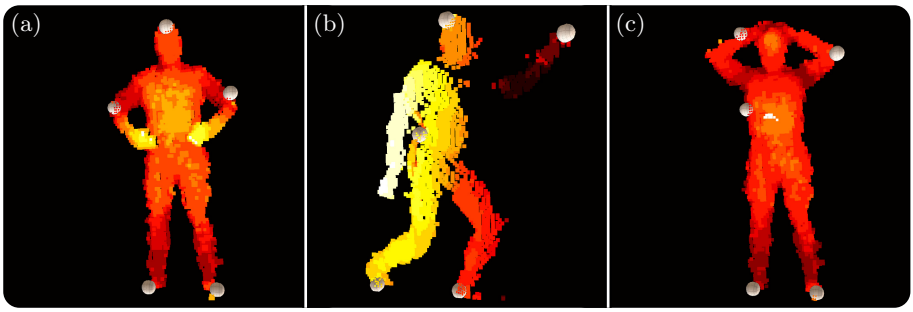

Fig. 9. Typical tracking situations when some of the geodesic extrema do not align with the hand, feet and head

to enable a densely sampled pose space while not requiring to sample the same pose in various global orientations. To this end, their database only contains poses, where the person is facing the camera frontally. As a consequence, also the query to the database needs to be normalized in the same way. By fitting a plane into a subset of depth pixels representing the torso of the person, the authors compute a front direction that serves as basis for the normalization, see also Fig. 8. Note that this way of normalization only works for near frontal poses and it is prone to noise and limbs occluding the torso. In order to pursue a normalization also in poses with occlusions, an additional inertial sensor could be leveraged to obtain a stable estimation of the person's front direction. This approach works for arbitrary rotations and is independent of optical clues that are prone to occlusions. This would already stabilize the lookup of poses from the database in cases when the geodesic extrema are calculated correctly.

However, there will be many occasions remaining where the query to the database, the geodesic extrema, cannot be calculated correctly. Some of these occasions with or without occlusions are shown in Fig. 9. The question is, whether it is possible to obtain poses from a database based on sparse features that are independent to occlusions. In computer animation this question is related to the data-driven reconstruction of human motions from sparse control signals. Many papers have come up that are inspired by an approach using sparse optical features presented in [41]. In particular, the two approaches [42,43] based on sparse inertial sensors data are interesting in our context since they do not rely on optical but inertial cues. In particular, the authors use the readings from inexpensive accelerometers fixed to the body to retrieve poses from a database. Unfortunately, the authors state, that using accelerometer data to obtain poses from the database is challenging because of the noisy characteristics of the data and the lack of discrimination of certain motions. This fact was further examined in [44], where the authors concluded that features based on orientations are better suited to describe full-body human motions than features based on accelerations. To conclude, a sparse set of inertial sensors could also be used to obtain a pose prior from a pose database when using *e. g.* orientation-based features are used for indexing. Such additional sensors could be easily added to the extremities of the person using straps.

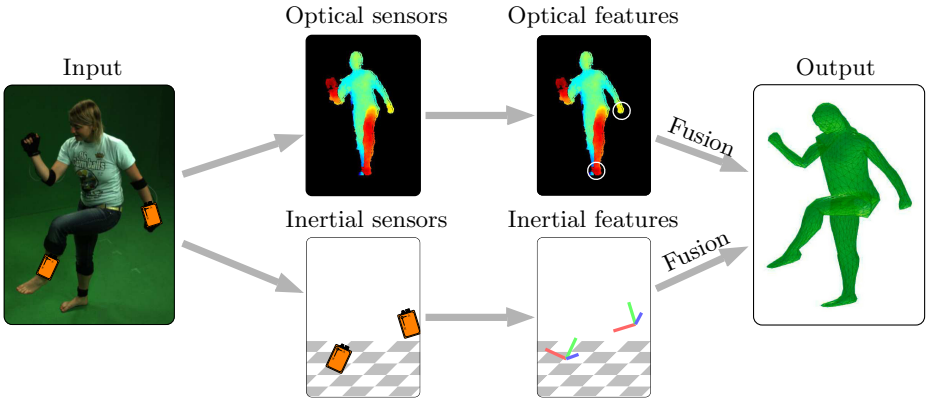

Fig. 10. Sketch of a fusion approach that uses optical depth data and inertial data to generate a single combined pose hypothesis.

Inertial data could also be used to support generative trackers. The idea is, to include information about the limbs orientations directly during the generative tracker's optimization. In contrast to the approach presented in [40], we propose not to solve two independent problems but building a combined energy function that incorporates visual and inertial constraints. In particular, optical cues might add positional constraints, while inertial sensors contribute with rotational constraints, see also Fig. 10. This would help to prevent tracking errors in a similar fashion as described in [7]. Furthermore, the inertial sensors would provide information about limbs even when they are not visible to the depth camera. This concept is modular in a way that one could selectively add inertial sensors to those parts of the body that need highly accurate tracking and do not attach sensors to body parts one does not need as accurate tracking. Overall, this enables selective tracking accuracy that can be adopted to the need of specific applications. Please note that the additional information needed to resolve rotational ambiguities might also be obtained from other sensor modalities such as RGB-input from a color camera. In particular, one could use feature tracking-based or optical-flow-based cues to stabilize tracking, see also [45].

4 Conclusion

In this chapter, we showed how recent depth cameras can be employed for tracking full-body human motion. Based on the unique properties of the provided depth data, such as easy background subtraction and geometric information, monocular tracking approaches become feasible that are not possible with traditional marker-less techniques. Furthermore, being much cheaper and easier to setup than systems used by traditional vision-based approaches, depth cameras, such as the Microsoft Kinect, have enabled applications even in uncontrolled

home user scenarios. While there was a lot of progress in the field of monocular depth tracking of human motions, current approaches still suffer from the challenging noise characteristics of depth cameras and the sparse information contained in their depth images. Especially rotational ambiguities and occlusions show, that the tracking of human poses is still very challenging and maybe not feasible in all cases when only relying to monocular depth images. To this end, we also discussed how current approaches could benefit from including additional, complementary sensor information for tracking stabilization. Here, work from other domains showed that inertial sensors are suitable to provide valuable information in cases when pure optical approaches fail.

References

1. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU 104(2), 90–126 (2006)
2. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: ICCV (2011)
3. Menache, A.: Understanding Motion Capture for Computer Animation and Video Games, 1st edn. Morgan Kaufmann Publishers Inc., San Francisco (1999)
4. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
5. Bregler, C., Malik, J., Pullen, K.: Twist based acquisition and tracking of animal and human kinematics. IJCV 56(3), 179–194 (2004)
6. Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR, pp. 1746–1753 (2009)
7. Pons-Moll, G., Baak, A., Helten, T., Müller, M., Seidel, H.P., Rosenhahn, B.: Multisensor-fusion for 3d full-body human motion capture. In: CVPR, pp. 663–670 (2010)
8. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: CVPR, pp. 1249–1256 (2011)
9. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: ICCV, pp. 951–958 (2011)
10. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR, vol. 2, pp. 126–133 (2000)
11. Starck, J., Hilton, A.: Spherical matching for temporal correspondence of non-rigid surfaces. In: ICCV, pp. 1387–1394 (2005)
12. Starck, J., Hilton, A.: Correspondence labelling for wide-timeframe free-form surface matching. In: ICCV, pp. 1–8 (2007)
13. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Computer Graphics and Applications 27(3), 21–31 (2007)
14. Matusik, W., Buehler, C., Raskar, R., Gortler, S., McMillan, L.: Image-based visual hulls. In: SIGGRAPH 2000, pp. 369–374 (2000)
15. de Aguiar, E., Stoll, C., Theobalt, C., Naveed, A., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. TOG 27, 1–10 (2008)
16. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. TOG (2008)

17. Pekelny, Y., Gotsman, C.: Articulated object reconstruction and markerless motion capture from depth video. CGF 27(2), 399–408 (2008)
18. Knoop, S., Vacek, S., Dillmann, R.: Fusion of 2D and 3D sensor data for articulated body tracking. Robotics and Autonomous Systems 57(3), 321–329 (2009)
19. Bleiweiss, A., Kutliroff, E., Eilat, G.: Markerless motion capture using a single depth sensor. In: SIGGRAPH ASIA Sketches (2009)
20. Friborg, R.M., Hauberg, S., Erleben, K.: GPU accelerated likelihoods for stereo-based articulated tracking. In: Kutulakos, K.N. (ed.) ECCV 2010 Workshops, Part II. LNCS, vol. 6554, pp. 359–371. Springer, Heidelberg (2012)
21. Demirdjian, D., Taycher, L., Shakhnarovich, G., Graumanand, K., Darrell, T.: Avoiding the streetlight effect: Tracking by exploring likelihood modes. In: ICCV, vol. 1, pp. 357–364 (2005)
22. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real-time human pose tracking from range data. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 738–751. Springer, Heidelberg (2012)
23. Ye, G., Liu, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Performance capture of interacting characters with handheld kinects. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 828–841. Springer, Heidelberg (2012)
24. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Realtime identification and localization of body parts from depth images. In: ICRA, Anchorage, Alaska, USA (2010)
25. Zhu, Y., Dariush, B., Fujimura, K.: Kinematic self retargeting: A framework for human pose estimation. CVIU 114(12), 1362–1375 (2010), Special issue on Time-of-Flight Camera Based Computer Vision
26. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: CVPR (2011)
27. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: ICCV, pp. 415–422 (2011)
28. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.W.: The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: CVPR (2012)
29. Salzmann, M., Urtasun, R.: Combining discriminative and generative methods for 3D deformable surface and articulated pose reconstruction. In: CVPR (2010)
30. Ganapathi, V., Plagemann, C., Thrun, S., Koller, D.: Real time motion capture using a single time-of-flight camera. In: CVPR (2010)
31. Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M.: Accurate 3d pose estimation from a single depth image. In: ICCV, pp. 731–738 (2011)
32. Liao, M., Zhang, Q., Wang, H., Yang, R., Gong, M.: Modeling deformable objects from a single depth camera. In: ICCV, pp. 167–174 (2009)
33. Krüger, B., Tautges, J., Weber, A., Zinke, A.: Fast local and global similarity searches in large motion capture databases. In: Symposium on Computer Animation, pp. 1–10 (2010)
34. Wei, X., Zhang, P., Chai, J.: Accurate realtime full-body motion capture using a single depth camera. TOG 31(6), 188:1–188:12 (2012)
35. Weiss, A., Hirshberg, D., Black, M.: Home 3D body scans from noisy image and range data. In: ICCV (2011)
36. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM TOG 24, 408–416 (2005)

37. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. CGF 2(28) (March 2009)
38. Maimone, A., Fuchs, H.: Reducing interference between multiple structured light depth sensors using motion. In: 2012 IEEE Virtual Reality Short Papers and Posters (VRW), pp. 51–54 (2012)
39. Butler, A., Izadi, S., Hilliges, O., Molyneaux, D., Hodges, S., Kim, D.: Shake'n'sense: Reducing interference for overlapping structured light depth cameras. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 1933–1936 (2012)
40. Ziegler, J., Kretzschmar, H., Stachniss, C., Grisetti, G., Burgard, W.: Accurate human motion capture in large areas by combining IMU- and laser-based people tracking. In: IROS, pp. 86–91 (2011)
41. Chai, J., Hodgins, J.K.: Performance animation from low-dimensional control signals. TOG 24(3), 686–696 (2005)
42. Slyper, R., Hodgins, J.K.: Action capture with accelerometers. In: Symposium on Computer Animation, pp. 193–199 (2008)
43. Tautges, J., Zinke, A., Krüger, B., Baumann, J., Weber, A., Helten, T., Müller, M., Seidel, H.P., Eberhardt, B.: Motion reconstruction using sparse accelerometer data. TOG 30(3), 18 (2011)
44. Helten, T., Müller, M., Tautges, J., Weber, A., Seidel, H.-P.: Towards cross-modal comparison of human motion data. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 61–70. Springer, Heidelberg (2011)
45. Brox, T., Rosenhahn, B., Gall, J., Cremers, D.: Combined region- and motion-based 3d tracking of rigid and articulated objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 402–415 (2010)