Tracking Humans Interacting with the Environment Using Efficient Hierarchical Sampling and Layered Observation Models

Jan Bandouch, Michael Beetz
Intelligent Autonomous Systems Group, Department of Informatics
Technische Universität München, Munich, Germany
bandouch@cs.tum.edu, beetz@cs.tum.edu
http://memoman.cs.tum.edu

Abstract

We present a markerless tracking system for unconstrained human motions which are typical for everyday manipulation tasks. Our system is capable of tracking a high-dimensional human model (51 DOF) without constricting the type of motion and the need for training sequences. The system reliably tracks humans that frequently interact with the environment, that manipulate objects, and that can be partially occluded by the environment.

*We describe and discuss two key components that substantially contribute to the accuracy and reliability of the system. First, a sophisticated hierarchical sampling strategy for recursive Bayesian estimation that combines partitioning with annealing strategies to enable efficient search in the presence of many local maxima. Second, a simple yet effective appearance model that allows for the combination of shape and appearance masks to implicitly deal with two cases of environmental occlusions by (1) subtracting dynamic non-human objects from the region of interest and (2) modeling objects (*e.g. *tables) that both occlude and can be occluded by human subjects. The appearance model is based on bit representations that makes our algorithm well suited for implementation on highly parallel hardware such as commodity GPUs.*

Extensive evaluations on the HumanEva2 benchmarks show the potential of our method when compared to state-of-the-art Bayesian techniques. Besides the HumanEva2 benchmarks, we present results on more challenging sequences, including table setting tasks in a kitchen environment and persons getting into and out of a car mock-up.

1. Introduction

Markerless human motion capture (HMC) has been in the focus of computer vision research for more than a decade now. The capability to observe articulated human motions on a joint level with unintrusive methods is much sought-after in areas such as human computer interaction (HCI), computer graphics and animation, or robotics, to name just a few. However, high-accuracy markerless motion capture systems to date are still constrained to well-defined and controlled environments (as evident in the case of the HUMANEVA [15] quasi-standard benchmarks), and are often constricted to predefined types of motion.

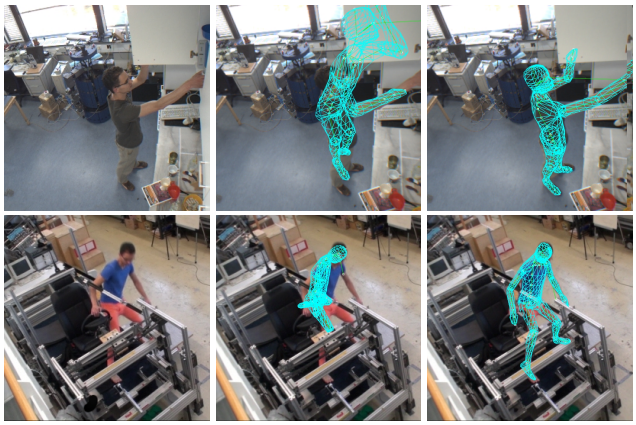

Figure 1. Two challenging setups for HMC featuring dynamic environments and occlusions. The center column shows the failure of shape-based methods. The right column shows the results using our appearance model with implicit environment modeling.

In our work we investigate how to apply markerless motion tracking techniques to activity observation tasks that include object manipulation and tasks that are performed in realistic unmodified (living) environments. The selected techniques must exhibit reliable performance in a large variety of scenarios with arbitrary types of motions while maintaining computational tractability. Additional challenges arising in this context are the segmentation of manipulated objects from the moving human and occlusion effects caused by environment objects that might also be dynamic as in the case of opening doors.

We present a working system that is capable of extracting high-dimensional human motion representations from challenging sequences (Fig. 1). Within this paper we will describe and discuss the techniques necessary to build such a system and motivate our choices. We will also intro-

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

duce novel methodological contributions that play a key role in making the system run reliably and with high accuracy: (1) We present a hierarchical sampling strategy that outperforms state-of-the-art particle-filter based techniques, especially when it comes to high-dimensional models. It combines partitioning and annealing strategies that make it highly efficient in overcoming local maxima of the weight functions. (2) We present a simple yet effective color-based appearance model that is computationally cheap and ideally suited for parallelization on commodity GPUs. (3) We show how to use this appearance model to not only enrich the observation model, but to provide a means of implicit environment modeling that can be used to disambiguate between human subjects, dynamic objects, and the environment in a simple and elegant way. Our representation does not require a 3D model of the environment, that is hard to come by in dynamic environments.

We provide extensive evaluation on the *HumanEva2* testset and on several other minute-long sequences for various application scenarios. Our main scenario is a kitchen environment, where humans are observed during everyday activities like cooking or setting the table. They frequently interact with objects (pick and place) and the environment (opening/closing cupboards and drawers), which leads to partial occlusions of the subjects. Further evaluation is provided on a sequence of a human getting into and out of a car mock-up, where strong occlusions occur.

The remainder is organized as follows. We briefly discuss related work in Section 2. Section 3 introduces our framework for human motion tracking, where we explain the novel hierarchical sampling strategy (3.1), the shape model (3.2), our efficient appearance model (3.3), and the layered representation for implicit environment modeling (3.4). To directly motivate our choices, we provide experimental evaluations along with the method descriptions in the corresponding subsections. We finish in Section 4 with a discussion of pros and cons of our method and conclude.

2. Related Work

In our work we are concerned with accurate markerless human motion capture from multiple cameras. Moeslund *et al*. [14] provide a survey and taxonomies of the field.

One of the most common image cues for pose estimation is shape, often encoded in silhouettes [10, 7, 3, 16] extracted using background subtraction. As silhouettes, especially in the monocular case, are ambiguous with respect to self occlusions, edges are often used to refine detection of human outlines [7, 3]. A richer description of shape can be obtained by calculating the visual hull from several camera views. Pose estimation is then performed by clustering rigidly-moving parts to estimate joint positions [6, 1], or by fitting a 3D model to the surface [12, 11]. While accurate, these methods require a large number of cameras to provide good results, and have not been tested under occlusions.

Color or texture information is less often used, partly due to the high computational cost. Balan and Black [2] use adaptive image templates that are updated over time. Kehl and van Gool [12] apply color information to visual hulls to find better nearest-neighbor correspondences between visual hull and model. Gall *et al*. [9] use an analysis-by-synthesis approach to detect point correspondences between the image and a textured surface model. This approach partly resembles our work with respect to the model synthesis, but is restricted to prominently textured clothing.

Pose tracking is usually performed from an initial estimate using either Bayesian approaches [7, 3] or deterministic optimization methods [12, 9]. Recently, exemplar-based methods have been in the focus of attention. These do not suffer from the high computational complexity in estimating the full DOFs of a human body. Two directions can be distinguished. A discriminative approach is to learn a direct mapping from observed image cues (such as silhouettes, SIFT, HOGs) to body poses [10, 16, 5], which can provide accurate estimates, especially given monocular vision. These methods are constricted to the type of features they have been trained with and training has to be redone when transfering to environments with different camera perspectives. Another exemplar-based approach is to learn a low-dimensional embedding of the human pose space for specific motions, that can be used to reduce the dimensionality of the problem [17]. However, such methods don't generalize over motions missing in the training set, and are difficult to extend to the observation of everyday activities.

The HUMANEVA benchmarks [15] have been adopted by the community as a means to compare different approaches [16, 5]. They consist of actors performing predefined actions like walking, running or boxing in a large and uncluttered environment. Little to none work is known to us where humans are observed performing everyday tasks in realistic environments, interacting with objects and the environment.

3. Human Tracking Framework

We perform model-based tracking of human motions in multi-camera environments with 3-4 static cameras. The model we use [4] is articulated through 51 DOFs and is able to provide realistic and highly accurate postures of human subjects (Fig. 2). It encodes biomechanical limits for joint angles as well as body-part dependant inter-frame variances constricting the amount of motion between consecutive frames. The outer shape of the model is represented through a surface triangle-mesh (< 2000 triangles) and can be adapted to different subjects.

The tracking problem is posed in a Bayesian framework as one of estimating the *posterior* probability density function (pdf) $p(x_t \,|\, y_{1:t})$ for the pose x_t at time t given a sequence of image observations $y_{1:t}$ up to time t. This *pdf* can be obtained recursively using *prediction* and *update* steps, given a *motion model* $p(x_t \,|\, x_{t-1})$ and an *observation model* $p(y_t \,|\, x_t)$. While it is computationally intractable to approximate the *pdf*, we use a hierarchical particle filter

framework to find its modes and thus the most likely pose estimate for each timestep.

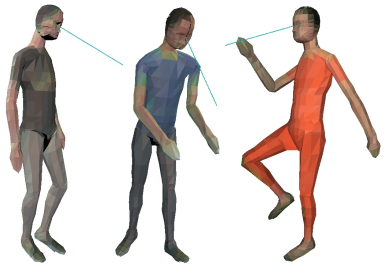

Figure 2. Three human model examples rendered in different poses with appearance information overlayed.

We currently expect that the first pose is initialized manually. This needs to be done only approximately due to the large convergence radius of the sampling strategy proposed in section 3.1. Experiments have shown convergence up to a pelvis translation of about $0.5\ m$ when the initial posture roughly resembles the observed pose.

Inner and outer model sizes also need to be manually set during initialization and remain constant throughout tracking. To simplify the model adaptation we performed a *PCA* of the size parameters from about 100 training exemplars. The adaptation using the principal components takes about 2 min for an experienced user.

In the next subsections we describe our hierarchical sampling strategy and the separate parts of the observation model in more detail. As for the motion model, we do not make use of informed motion priors to make sure we are able to track any posture that can be physiologically described by our model, which is an important prerequisite in the observation of everyday tasks. Thus, we simply propagate the last particle state and diffuse it with Gaussian noise according to the body-part dependant inter-frame variances.

3.1. Hierarchical Sampling Strategy

Due to the non-linearity in both motion and observation models, particle filters [8] are a common choice for estimating the *pdf*. However, standard *Sampling Importance Resampling* (SIR) fails to approximate high-dimensional *pdf*s due to the exponential growth in particles needed. Therefore, one usually tries to estimate only the modes of the *pdf*. Two sampling strategies have been proposed that claim to do this efficiently, namely the *annealed particle filter* (APF) [7] and *partitioned sampling* (PS) [13]. In APF, particles are gradually moved towards the global maximum of the weight function in several iterations (layers). By bluring the weight function in the initial layers, particles are able to escape local maxima in early iterations, similar to simulated annealing. PS on the other hand provides a hard partitioning on the state space, where the resulting smaller partitions are estimated in sequential order using SIR.

While APF has been shown to provide good results in HMC on some short sequences, PS has only been proposed for hand tracking. Bandouch *et al*. [4] recently evaluated both approaches in the context of HMC with high-dimensional models, and proposed a combination of both, hinting at their complementary strengths. The intuition is that APF as a soft partitioning approach still suffers from exponential growth, whereas PS suffers from bad estimates in early partitions. By using annealing layers inside the partitions, the size of these partitions can be increased, resulting in better overall estimates. However, it remains unclear what type of hierarchical partitioning and also what order of partitioning (*e.g*. arms first or legs first?) is best. While this might not make a difference for noise-free simulation data, in practice an incorrectly estimated early partition due to noisy data irrevocably misleads the outcome (Fig. 4).

We ran a series of experiments on the *HumanEva2* benchmark to verify the assertions from [4] and to find an improved sampling strategy that is highly reliable in practice given noisy observations. To provide comparable results, all experiments were performed using our human model (38 DOF; ignoring lower spine, hands and feet) in combination with the shape-based observation model described in section 3.2. We first evaluated the APF (Fig. 3a), testing three different variants for constricting the amount of diffusion to each body part with each new layer. In the original version, the diffusion is reduced using a constant factor. The first variant controls the amount of diffusion according to the variances of particle states from the last iteration. In the second variant, particles are diffused using the state covariance matrix from the particle set in the last iteration. This covariance scaled diffusion also provided the best results, although the overall performance of the APF is unsatisfying. While the torso is correctly estimated most of the time, the limb positions are often wrong. We attribute this to the fact that all body parts are estimated at once, which results in a random shuffling of the outer limbs until their preceding body parts could be localized.

We then tested PS (Fig. 3b) at comparable processing times (1600 particles PS $\sim$ 10 layers 800 particles APF). The order of partitioning was pose and lower torso, upper torso, left thigh, left lower leg, right thigh, right lower leg, left arm, left forearm and so on. Again, the results are unsatisfying, which can be attributed to the large ambiguity in the weight function when evaluating the lower torso.

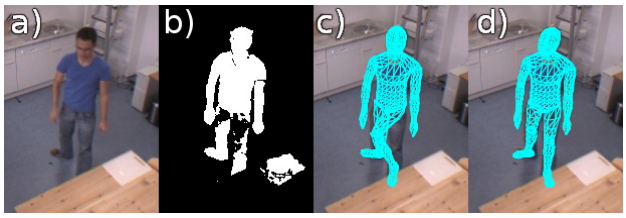

Figure 4. The order of limb partitioning can influence the outcome given noisy observations: a) Original image b) Noisy foreground mask c) Left leg first partitioning d) Right leg first partitioning

During the experiments we observed that APF manages to find good estimates of the torso, but fails in estimating

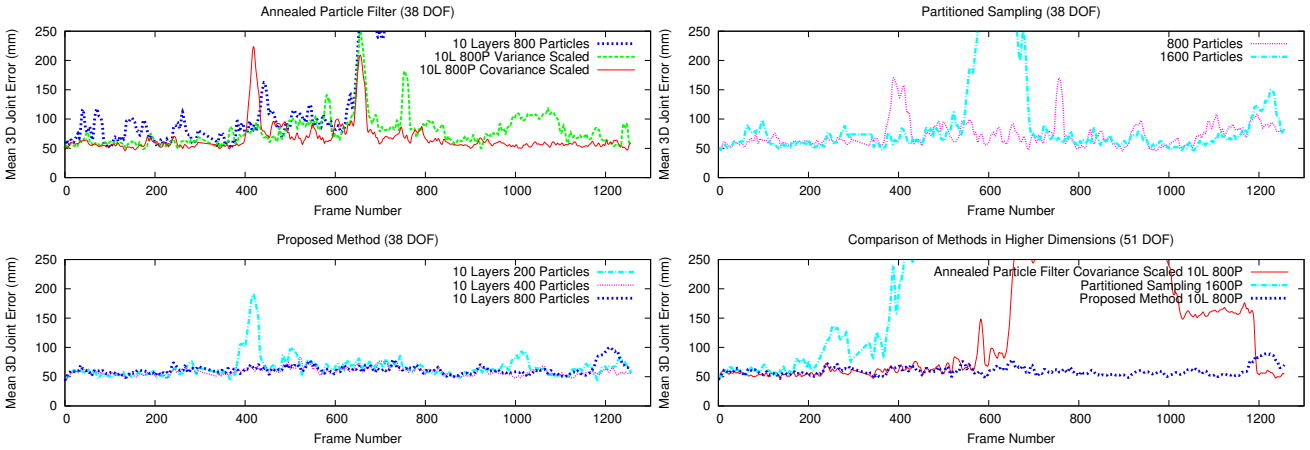

Figure 3. Tracking results on the *HumanEva2* benchmark (sequence S4). Top-left: Annealed Particle Filter (38 DOF). Top-right: Partitioned Sampling (38 DOF). Bottom-left: Proposed Method (38 DOF). Bottom-right: Comparison of all methods (51 DOF). Note that there seems to be a systematic error (the error never drops below 5 cm) resulting from differences in relative joint positions between the ground truth model and ours. Visually, our proposed method delivers near perfect results (see `http://memoman.cs.tum.edu` for videos).

the limbs. PS on the other hand is bad at estimating the torso, but is able to estimate limb positions accurately when given a good torso estimate. Thus, a combination of APF for estimating the torso with PS for estimating the limbs turns out to be beneficial for tracking. However, PS suffers from another problem in practice. Whenever dealing with noisy observations, the order of partitioning influences the local maxima encountered on the way, which often results in limbs being interchanged (Fig. 4). This effect can be weakened by introducing annealing steps to each partition, but the overall problem remains. To overcome this problem, we propose to divide particle evaluations into parallel pipelines of partitions with different sizes and order of estimation (*e.g.* left arm first then right arm or vice versa or both arms in one partition with annealing). Afterwards, particles emerging from the different pipelines are combined and reweighted, so that the best strategy wins, as in a voting scheme. The high diversity of local sampling strategies turns out to be the key to robust behavior by avoiding to get stuck in local minima created by noisy image observations.

Our newly proposed sampling strategy consists of initial covariance scaled annealing steps, where we estimate only the state of the torso. Annealing proceeds until the iteration terminates or the variances from the last iteration drop below a threshold. Then, several parallel sampling strategies for the limbs are started with different partition sizes and order of execution. These also involve annealing steps, but using less layers. Each of the partitions only uses a fraction of the original particles, depending on the size of the partition and on the number of parallel partitions. Finally, particles from the parallel partitions are resampled based on their weights, leaving the fittest to survive.

Due to space constraints, we can give only marginal comments on implementation details of our method. Annealing partitions are implemented in the same way as described in [7], while the resampling of particles from one partition to another is performed as described in [13]. Parallel partitions are resampled into a new partition by concatenating the previously split particle sets and updating the weights for the concatenated set before proceeding with the next resampling step. We will not comment on the exact size and order of the parallel partitions here, but we expect the general concept to work well with different choices of partitions, as long as enough diversity in partitioning is ensured (changing order of limb evaluations, changing size of partitions). As a rule of thumb, the bigger the partitions become, the more annealing layers should be introduced.

The results of our proposed approach on the benchmark (Fig. 3c) show stable behavior and clearly outperform both APF and PS. Even when using only a fraction of overall particles (200 instead of 800), results are still much better than in the former approaches. What is more, we repeated the experiments using the best variants of each algorithm, this time tracking the full 51 DOF of our model, including the lower spine, hands and feet. Both APF and PS get completely confused, while our strategy provides almost the same quality results as in the 38 DOF case, due to the hard partitioning of the state space (Fig. 3d). To the best of our knowledge, our sampling strategy is the first to have shown tracking success with such high-dimensional models using a Bayesian approach without making use of training data.

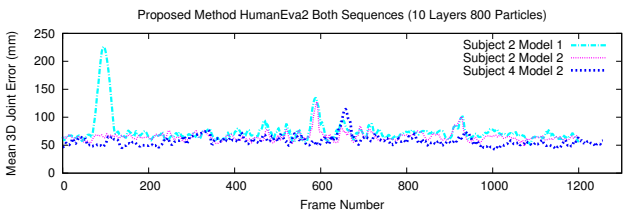

Figure 5. Results of the proposed method on all HumanEva2 sequences using different models than in Fig. 3

We also provide results of our approach on the other sequence in the *HumanEva2* benchmark, using different

model adaptations (Fig. 5). As an example, model 1 for sequence 2 was too small, resulting in slightly worse behavior. Note that we used exactly the same parameters for all experiments in this paper, and we did not learn or tweak parameters towards specific sequences. This especially means that although *HumanEva2* has been recorded at 60 Hz, we used the larger inter-frame variances suitable for our own 25 Hz recordings. Processing took about 20 sec per frame using our single-threaded C++ implementation on a consumer laptop.

3.2. Shape Model

We use a very simple and common observation model inside the particle filter framework, that is based on silhouette shapes extracted from multiple cameras. Besides its simplicity, it has several advantages over more complex shape models such as visual hulls when it comes to occlusions and dynamic objects, as will be pointed out later.

For better clarity we introduce the following notations for logical operations on binary image masks I:

$$\begin{aligned}
\mathrm{AND}(I_A, I_B)[x,y] &= I_A[x,y] \wedge I_B[x,y] \\
\mathrm{OR}(I_A, I_B)[x,y] &= I_A[x,y] \vee I_B[x,y] \\
\mathrm{DIFF}(I_A, I_B)[x,y] &= I_A[x,y] \ominus I_B[x,y] \\
\mathrm{XOR}(I_A, I_B)[x,y] &= I_A[x,y] \triangle I_B[x,y] \\
\mathrm{NOT}(I)[x,y] &= \neg I[x,y] \\
\mathrm{COUNT}(I) &= \sum_{x,y} I[x,y]\,; \qquad I[x,y] \in \{0,1\}
\end{aligned}$$

Here, $I[x,y]$ corresponds to the binary value of the image mask I at pixel $[x,y]$. The four binary operators AND, OR, DIFF and XOR correspond to pixelwise intersection ($\wedge$), union ($\vee$), difference ($\ominus$) and symmetric difference ($\triangle$) operations on the two image masks I_A and I_B. NOT specifies inversion ($\neg$). To simplify matters, we assume that these operations work on the image masks of all cameras in parallel. The COUNT operation then sums up the non-zero pixels of all cameras.

For evaluating the quality of a pose estimate (*i.e.* particle), the shape of human subjects is evaluated by silhouette comparison of the model projection masks I_P and binary foreground masks I_F extracted using a background subtraction technique. In this notation, the commonly proposed SSD error measure [3, 7] between selected points on the model and the foreground masks would correspond to $\mathrm{COUNT}(\mathrm{DIFF}(I_P, I_F))$. However, we use the sum of a pixelwise logical symmetric difference as error measure e_S:

$$e_S = \mathrm{COUNT}(\mathrm{XOR}(I_P, I_F)) \tag{1}$$

This is justified in that the projection should not only be explained well by the foreground mask, but also should the foreground mask be covered as good as possible by the projection. When ignoring this, arms tend to stick to the torso unless other visual cues, *e.g.* edges (that tend to be unreliable with clothing), are incorporated. However, it should be noted that using the symmetric difference requires the whole image mask to be processed on the contrary. We alleviate the increased computational expense by representing image masks using runlength-encoding.

3.3. Appearance Model

When using three or more cameras, silhouette shapes are a sufficiently rich and unambiguous descriptor of human poses [3, 4], as also proved in our experiments. This changes in the presence of additional foreground objects, or when there are occlusions from the environment. Color-based appearance models can help to disambiguate the foreground, but the large computational expense of such methods, that often involves calculation of Mahalanobis distances to given color clusters (for each pixel and for every particle), makes them tedious to use in practice.

We introduce a simple yet effective appearance model where color channels are represented as bitmasks. The 3D colorspace is divided into voxels of equal size. Each set bit in a bitmask represents the index of color voxels inside the color cluster. When using 32 bit Integers for each color channel, the colorspace is divided into a grid of $32 \times 32 \times 32 = 32768$ different colors, and a color cluster is represented by 3 integers. Testing whether a color is inside a given color cluster now becomes very efficient using bitwise logical AND operations, and correspondence is established when all three resulting Integers are non-zero. This computation can be highly optimized using techniques such as loop-unrolling, or parallelization on GPUs.

Unfortunately, in this representation it is only possible to represent rectangular clusters of colors in the 3D colorspace. To account for the typical shape of Gaussian color distributions, where the principal component is aligned along the luminance direction, we thus propose to use the *HLS* colorspace, which better captures the shape of typical color distributions in man-made environments (Fig. 6). To account for the fact that hue and saturation loose significance at high and low luminance, and that hue also looses significance at low saturation, additional masking is applied to the bitmasks of respective colors using precomputed lookup-tables.

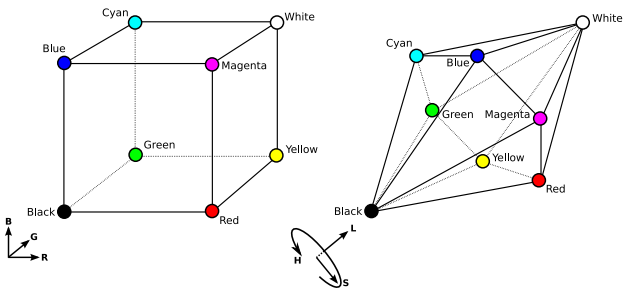

Figure 6. Left: RGB color cube; Right: HLS dual cone color representation at comparable spatial alignment.

Color clusters for each triangle in the human mesh are initialized from the given first pose, and are refined during tracking. By rendering our model using a z-buffer algorithm with the triangle index as color, we can associate visible colors to the triangles in our model. We only add new colors when they are sufficiently close to the cluster mean, and when the triangle normal is approximately facing towards a camera. The corresponding bitmask representations are calculated considering the mean and the eigenvalues of the colors in each cluster. In the future we plan to shift the luminance mean in each frame according to a histogram comparison between frames to account for changes in lighting.

We incorporate the appearance model into our observation model by estimating how well a rendered color projection of our model matches with the image colors. The sum of all consistent color values forms the correlation measure c_A. Given the error e_S from our shape model, a combined weight ω is calculated as follows:

$$\omega = \alpha \cdot (1 - \mathrm{NORM}(e_S)) + \beta \cdot \mathrm{NORM}(c_A) \quad (2)$$

where α and β are constants used for balancing the contributions of shape and appearance (we use $\alpha = 0.7$ and $\beta = 0.3$), and NORM scales the error and correlation measures to an interval between 0 (lowest encountered value) and 1 (highest encountered value).

We have implemented the combined observation model on a GPU using the NVidia CUDA API, and were able to get a speed gain of $\sim$2 when compared to the optimized CPU implementation of only the shape model using runlength-encoding. Experiments on the quality of tracking indicate that color-based tracking is beneficial when estimating the head orientation, but does not necessarily improve the overall tracking quality of other body parts, or reduce the amount of particles needed.

3.4. Implicit Environment Model

The learned appearance for our human model can be used to cope with two problematic cases encountered in real environments, namely the presence of dynamic non-human foreground objects, and occlusions by the environment. For dealing with these cases, we introduce a new binary layer mask I_B into our observation model, that will be used to block regions from processing. This mask is set (1) for all pixels that should be processed, and unset (0) for regions to be blocked. Blocking is then achieved by masking out the respective parts in both the foreground mask I_F and the projection mask I_P before evaluating the shape error e_S and the appearance correlation c_A:

$$I_F = \mathrm{AND}(I_F, I_B) \quad (3)$$
$$I_P = \mathrm{AND}(I_P, I_B) \quad (4)$$

Dynamic non-human foreground objects: In this case dynamic objects (possibly manipulated by the human subject) or dynamic parts of the environment (doors, cupboards, drawers) appear inside the foreground mask and mix with the human silhouettes. To filter these parts, we introduce a human appearance mask I_H that is set whenever a pixel's color resembles a color in our appearance model. Such a mask can be calculated efficiently using a binary lookup table for each color voxel index (32768 entries) that is calculated once when the appearance model is updated. We then remove non-human parts from the foreground mask I_F and add them to the blocking mask I_B using the following operations:

$$I'_F = \mathrm{AND}(I_F, I_H) \quad (5)$$
$$I'_B = \mathrm{AND}(I_B, \mathrm{NOT}(\mathrm{DIFF}(I_F, I'_F))) \quad (6)$$

Adding non-human parts to the blocking region is important, as we have no idea whether the dynamic object might be occluding the human. It also weakens the influence of erroneously removed parts of the humans, as they will be ignored without penalizing the shape model. Fig. 7 gives an example on the effectivity of these simple operations.

Environmental occlusions: In this case static objects in the environment partially occlude the observed subjects (*e.g.* tables, chairs). However, these objects can also be occluded by the subjects, as there is no persistent spatial ordering. We mark regions that are candidates for occlusions (*e.g.* tables) by unsetting these regions in the blocking mask I_B. This needs to be done once during camera setup and can be done by a user within seconds by choosing a polygonal region to be considered. Such regions will by default be ignored during evaluation. To prevent valid observations of human body parts to be blocked, *e.g.* when arms are visible above a table, we exclude all human-like foreground regions from blocking:

$$I''_B = \mathrm{OR}(I'_B, I'_F) \quad (7)$$

Fig. 8 shows exemplar results using this kind of occlusion modeling. The amount of occlusion that can be compensated depends on the number of cameras used and their placement, *i.e.* each body part should always be observable from at least three cameras. Scenarios with more occlusion thus require more cameras.

Experimental Results: We have evaluated our system including the implicit environment models on several minute-length sequences in a kitchen environment, where humans are observed while setting a table. Although we are not able to provide quantitative results on these sequences due to missing ground truth data, videos are available at `http://memoman.cs.tum.edu`. In these sequences, several actors (with different body shapes) frequently interact with the environment by picking up objects from drawers or cupboards and placing them on a table. Furthermore, while standing near the table, the legs are occluded in two out of four views. When using only the shape-based observation model, the results of our tracker are either inaccurate

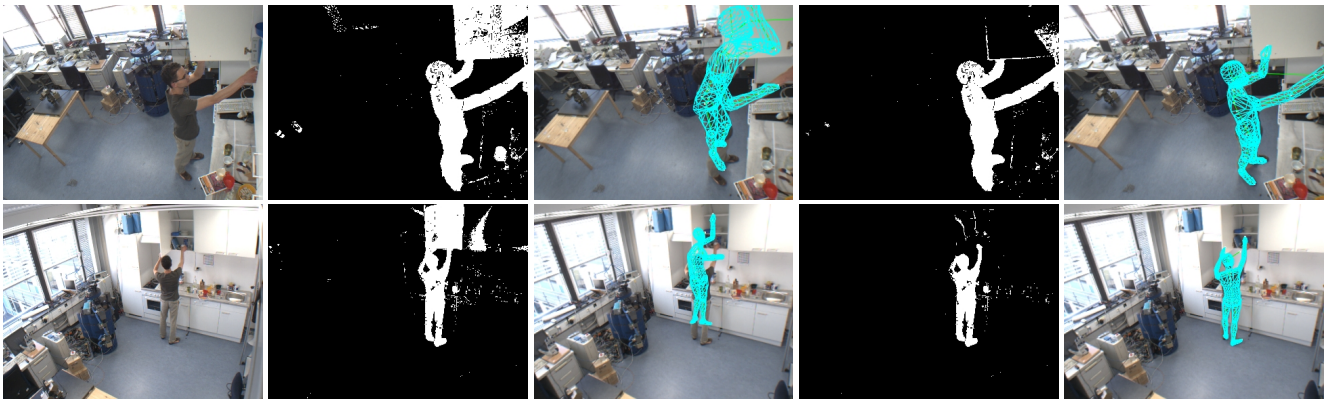

Figure 7. Example frame with opened cupboard (from left to right): a) original image b) unmodified foreground mask I_F c) tracking results without using appearance d) foreground mask I'_F after non-human foreground removal e) tracking results using our method. Each row shows one camera view. White color in the mask represents set bits.

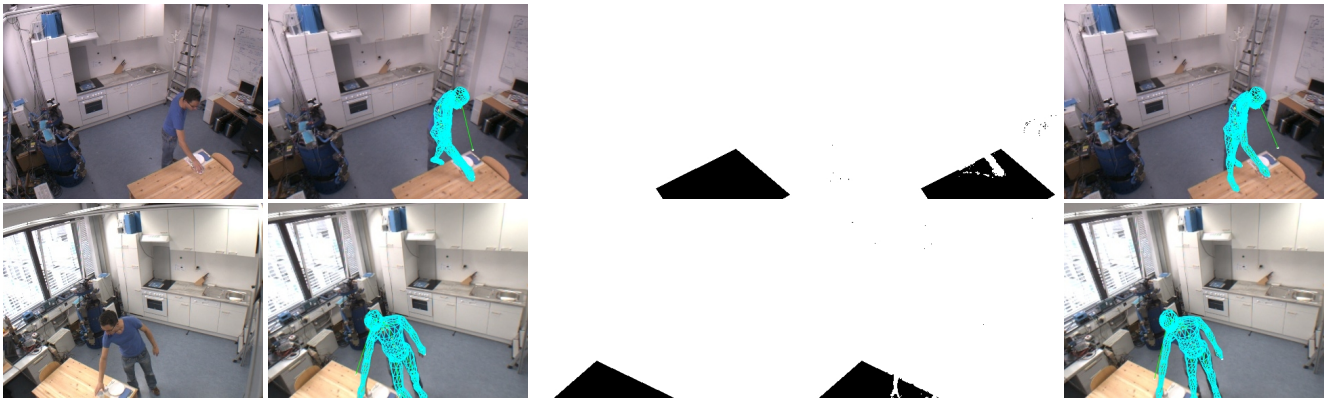

Figure 8. Example frame with occlusion from table (from left to right): a) original image b) tracking results without using blocking layer and appearance c) original user-specified blocking mask I_B d) blocking mask I''_B after exclusion of human-like parts e) tracking results using our method. Each row shows one camera view. White color in the mask represents set bits (blocked regions are black).

or fail completely (Fig. 7 and Fig. 8). Using the implicit environment modeling, the sequences can be correctly processed without the need for reinitialization (Fig. 9). Most of the problems encountered in our evaluation either stem from the failure to separate peoples silhouettes from the background, or from inaccurate body models, which indicates the importance of good body models in accurate HMC.

We have further evaluated our methods in a completely different setup, where we track persons while getting into and out of a car mock-up as used in ergonomic studies. Here, occlusions given by the mock-up and the seat are much stronger than in the kitchen sequences. Nonetheless, we were able to track a sequence of 2500 frames at good accuracy (Fig. 10). The original video sequences to our experiments are available on request.

4. Discussion and Conclusion

We have shown how to reliably track human fullbody motions with high accuracy for challenging tasks such as everyday manipulation activities involving dynamic objects and frequent occlusions in realistic environments. A key aspect herein is our robust sampling strategy that is able to find yet unobserved states. Without the ability to sample relevant parts of the state space, the best observation models will fail. While our observation models seem simple, they are computationally tractable, provide good accuracy, and are well-suited for the kind of implicit environment modeling we propose. To achieve comparable results for the sequences presented using *e.g.* visual hulls, the amount of cameras necessary would probably be untractable.

Although there are still steps that need manual initialization, the necessary amount of user input both for pre- and post-processing is small when compared to marker-based tracking systems. Our ongoing research is aimed at automating all initialization steps and at developing improved motion models for efficient prediction of dynamics. This will help to reduce the number of particles needed for tracking. Further speed-ups are expected from fully exploiting the potential for parallelization inherent to our particle-filter-based framework.

References

[1] D. Anguelov, D. Koller, H.-C. Pang, P. Srinivasan, and S. Thrun. Recovering articulated object models from 3d range data. In *AUAI*, 2004.

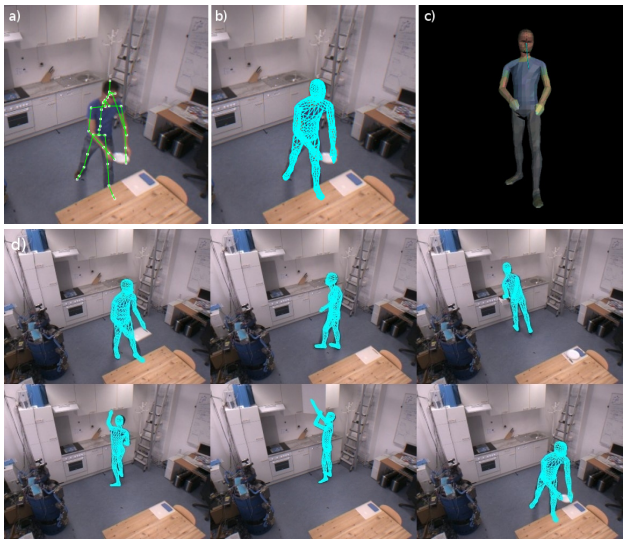

Figure 9. Kitchen sequence (one out of four views): a) inner model b) outer model c) virtual 3D view with appearance model d) screenshots from the sequence (1300 frames or 52 sec). Notice the interaction with objects, drawers and cupboards.

[2] A. O. Balan and M. J. Black. An adaptive appearance model approach for model-based articulated object tracking. In *CVPR*, 2006.

[3] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *ICCCN*, 2005.

[4] J. Bandouch, F. Engstler, and M. Beetz. Evaluation of hierarchical sampling strategies in 3d human pose estimation. In *BMVC*, 2008.

[5] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3d prediction. In *CVPR*, 2008.

[6] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *CVPR*, 2003.

[7] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV)*, 61(2):185–205, 2005.

[8] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[9] J. Gall, B. Rosenhahn, and H. Seidel. Drift-free tracking of rigid and articulated objects. In *CVPR*, 2008.

[10] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, 2003.

[11] R. P. Horaud, M. Niskanen, G. Dewaele, and E. Boyer. Human motion tracking by registering an articulated surface to 3-d points and normals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[12] R. Kehl and L. V. Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding (CVIU)*, 104(2):190–209, 2006.

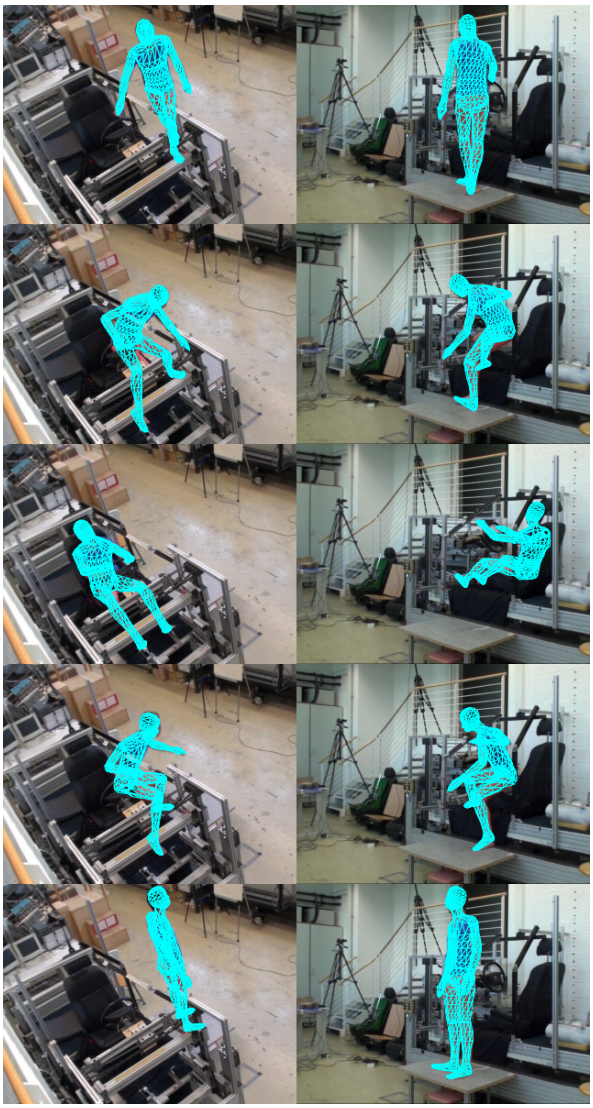

Figure 10. Screenshots from the mock-up sequence (2500 frames or 100 sec). Two out of four cameras are shown in the columns. Notice the strong environmental occlusions from bars and the seat.

[13] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV*, 2000.

[14] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2):90–126, 2006.

[15] L. Sigal, L. Sigal, M. J. Black, and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, 2006.

[16] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *CVPR*, 2008.

[17] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006.