Tracking People Interacting with Objects

Hedvig Kjellström[1] Danica Kragić[1] Michael J. Black[2]

[1] CVAP/CAS, CSC, KTH, 114 22 Stockholm, Sweden, {hedvig,danik}@kth.se

[2] Dept. Computer Science, Brown University, Providence, RI 02912, USA, black@cs.brown.edu

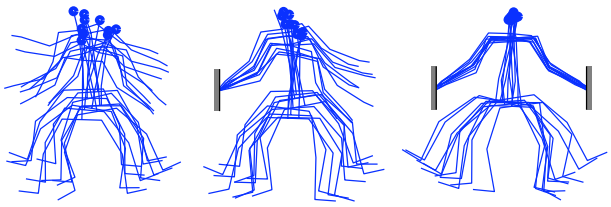
Abstract

While the problem of tracking 3D human motion has been widely studied, most approaches have assumed that the person is isolated and not interacting with the environment. Environmental constraints, however, can greatly constrain and simplify the tracking problem. The most studied constraints involve gravity and contact with the ground plane. We go further to consider interaction with objects in the environment. In many cases, tracking rigid environmental objects is simpler than tracking high-dimensional human motion. When a human is in contact with objects in the world, their poses constrain the pose of body, essentially removing degrees of freedom. Thus what would appear to be a harder problem, combining object and human tracking, is actually simpler. We use a standard formulation of the body tracking problem but add an explicit model of contact with objects. We find that constraints from the world make it possible to track complex articulated human motion in 3D from a monocular camera.

1. Introduction

An overwhelming majority of human activities are interactive in the sense that they relate to the scene around the human. For example, people are supported by the floor, chairs, ladders, etc., they avoid obstacles, and they push, pull and grasp objects. This suggests that tracking and recognition of human motion from images would benefit from employing contextual information about the scene including the objects in it as well as other humans. Despite this, visual analysis of human activity has rarely taken scene context into account [16]. Algorithms for 3D human tracking, in particular, often view the human body in isolation, ignoring even the common interactions with the ground plane. We argue that, not only is it important to model interactions between humans and objects, but it makes tracking human motion *easier*.

Without loss of generality, we focus on the case of monocular tracking of a human interacting with a single rigid object in the scene. 3D human tracking – especially from a single view – is understood to be a difficult problem due to the high-dimensional, non-rigid, and articulated structure of the human body. Object detection and tracking, in contrast is often much simpler. Recent advances in object recognition show that, for wide classes of rigid objects, reliable detectors can be built. Moreover, for known rigid objects, good solutions exist for real-time pose estimation and tracking. If objects can be detected and tracked, we argue that this can simplify the human tracking problem.



(a) (b) (c)

Figure 1. Contact with objects of known pose constrains the space of possible body poses. (a) Unconstrained. (b) One hand constrained. (c) Both hands constrained. This figure shows the principles behind our approach, using a 2D model for visibility. The actual model used in the experiments is defined in 3D.

Specifically we consider what happens when a human holds a rigid object in their hand. If the pose of the object can be determined, along with some details of the grasp, then the object pose can constrain the pose of the hand. This constraint then propagates along the kinematic body tree to limit the possible range of body poses. The idea is summarized in Figure 1. The key insight is that we treat objects as "extra body parts" that may, or may not, be attached to the body at a given time. When attached, the pose of the object and the relative degrees of freedom between the object and the human hand constrain the possible body poses.

Consider the problem of estimating the ulnar-radial rotation of the forearm in video; often this degree of freedom is poorly constrained by image observations. If the person now grasps a baseball bat, the pose of the bat greatly constrains the possible forearm rotation. Unlike previous models, our state space includes, not only the joints of the body,

but contact relationships between body parts and scene objects. Of course different types of contact are possible including various types of attachment which imply different kinematic constraints. Here we focus on two simple cases: 1) contact proximity, in which one or both hands are attached to an object, constraining their distance from it but not the relative orientation; 2) interpenetration, in which the body is constrained to not pass through other known objects. Our implementation extends a standard kinematic 3D body tracker using annealed particle filtering for optimization [8].

We demonstrate that information about object pose and contact can be incorporated in a kinematic tree body tracker and that doing so improves the accuracy of body tracking in challenging situations. The key observation is that object tracking is often easier than body tracking. By explicitly formulating body-object contact relationships in the state space, we can infer contact and constrain the body kinematics. Viewed another way, grasped objects can be thought of as "natural markers" that can be tracked to simplify the body pose estimation problem. We test our method in the case of single-view 3D reconstruction of complex articulated human motion without priors.

2. Related Work

Tracking and reconstruction of articulated human motion in 3D is a challenging problem that has been investigated thoroughly during the last decade [1, 2, 7, 8, 17, 19, 21, 22, 24]. Given the difficulty of tracking high-dimensional body structure, present methods all impose limitations to constrain the optimization, either in their assumptions about image appearance (e.g. static background), the imaging environment (e.g. multiple calibrated cameras), the range of allowed motions, or in the number of degrees of freedom in the human body model. Here we explore different constraints from objects in the environment.

Common to almost all human tracking methods is that the human is considered in isolation. There has been work on using human movement to recognize objects (e.g. [9]) but less on using object recognition to constrain the interpretation of human movement. Notable exceptions are methods that exploit human-scene context for recognition of human activity involving objects [10, 12, 15, 13]. There, the human and the scene are related on a semantic level, in that the object type constrains the activity and vice versa.

Several methods have explored human interactions with the environment and how these affect human appearance and movement. Bandouch and Beetz [3] model occlusion of the human from complex objects in the scene. Others focus on contact with the ground plane [26, 27] and constraints on human motion from the spatial layout of the scene [14]. Yamamoto and Yagishita [26] in particular propose an incremental 3D body tracking method that also constrains the position, velocity and acceleration of the body using scene constraints. They consider several highly constrained situations, such as constraining human arm motion given the known trajectory of a door knob and the assumption of contact, or tracking a skier with constraints on foot position, orientation and velocity relative to the known ski slope. We go beyond their work to actually track the movement of an object in the scene and to infer the contact relationships between the human and the object. These are critical steps for making more general scene constraints practical.

Dynamical models of human activity have also been used to incorporate interactions of the body with the world (again mostly the ground plane) [5, 6, 25]. The approach of Vondrak *et al.* [25] is promising because they exploit a dynamics simulator that acts as a prior on human movement. If environmental structure is available, this general approach can simulate interactions between the body and the world. Our approach differs in that we focus on constraining kinematics rather than dynamics.

We propose to aid human tracking by taking objects in the human's hands into regard. A first step in this direction is taken by Urtasun *et al.* [23] who address the special case of tracking a golfer's hands by robustly tracking the club through the swing. They note, as do we, that the problem of rigid-object motion reconstruction is vastly easier than that of reconstructing human motion. Their work focuses on the case of monocular tracking of a stylized movement for which they have a strong model. Specifically they use the position of the golf club to constrain the temporal location of the pose over the swing movement. A similar idea is exploited by Gupta *et al.* [11] who employ contextual constraints in motion reconstruction, such as the speed and direction of a ball in tennis forehand reconstruction, or the height of a chair in reconstruction of a sitting action.

The difference between both these approaches and our method is in the way the contextual object information is exploited. While they employ the object observations in a discriminative manner, we explicitly take object pose into account in a generative model of human pose. They also exploit strong prior models of temporal movement and model how the external environment interacts with these models. In contrast, we forgo the temporal prior; this allows us to track previously unseen movements.

Similarly to us, Rosenhahn *et al.* [18] model known kinematic constraints from objects in the world. They use a region-based tracker which allows for a compact analytic formulation of the constraints; we instead use a numerical method. The advantage of our kinematic chain tracker is its lower number of degrees of freedom, which allows for tracking with less image information.

3. Articulated Tracking

To study the effect of object constraints we use a simple, standard and well understood tracking framework which we

then modify. The observations, of course, apply to more sophisticated tracking frameworks as well.

The human is modeled as an assembly of connected truncated cones representing different limbs. The articulated pose of the human model at time t is defined by 40 parameters, $\boldsymbol{\alpha}_t$; these include pose with respect to a global coordinate system, and 34 Euler angles defining the relative pose of the limbs. From $\boldsymbol{\alpha}_t$, it is possible to derive T_t^{gl}, the transformation matrix from the global coordinate system to the local coordinate system of each limb l. This makes it possible to determine the position of all surface points on the human model in the global coordinate system as $\mathbf{p}_t^g = T_t^{gl}\mathbf{p}^l$.

3D articulated human tracking is a problem of time-incremental search in a very high-dimensional state space $\mathcal{S}_\alpha$. It is well studied (Section 2), and several different methods exist. We here use annealed particle filtering (APF) [8], which searches for the mode of a probability density function over pose $\boldsymbol{\alpha}_t$ given the history of image observations $\mathbf{D}_{1:t}$ represented as:

$$p(\boldsymbol{\alpha}_t \,|\, \mathbf{D}_{1:t}) \propto$$
$$p(\mathbf{D}_t \,|\, \boldsymbol{\alpha}_t) \int_{\mathcal{S}_\alpha} p(\boldsymbol{\alpha}_t \,|\, \boldsymbol{\alpha}_{t-1}) p(\boldsymbol{\alpha}_{t-1} \,|\, \mathbf{D}_{1:t-1}) d\boldsymbol{\alpha}_t \; . \quad (1)$$

The posterior model $p(\boldsymbol{\alpha}_{t-1} \,|\, \mathbf{D}_{1:t-1})$ is crudely represented by a collection of N samples, or particles $\boldsymbol{\alpha}_{t-1}^i$, each with a weight w_{t-1}^i. At a given timestep, N new particles are sampled from the previous weighted set using Monte Carlo sampling to produce a new set of unweighted particles $\{\tilde{\boldsymbol{\alpha}}_{t-1}^i\}_{i=1}^N$. Each particle is then propagated in time by sampling from the temporal update model $p(\boldsymbol{\alpha}_t \,|\, \tilde{\boldsymbol{\alpha}}_{t-1}^i)$, producing a new unweighted set $\{\boldsymbol{\alpha}_t^i\}_{i=1}^N$. Each particle is then evaluated against the observed data $\mathbf{D}_t$, to compute a weight using the likelihood model $w_t^i = p(\mathbf{D}_t \,|\, \boldsymbol{\alpha}_t^i)$. The weights are normalized to sum to 1.

In each timestep, the APF weights and resamples the posterior distribution A times using smoothed versions of the likelihood distribution, in a simulated annealing manner (hence the name). This is further described below.

3.1. Temporal Update Model

$p(\boldsymbol{\alpha}_t \,|\, \boldsymbol{\alpha}_{t-1})$ is defined using a zero-order linear motion model with Gaussian noise with covariance C:

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\nu} \; , \; \boldsymbol{\nu} \sim \mathcal{N}(0, C) \; . \quad (2)$$

The covariance matrix C is learned from the motion capture data in the HumanEva dataset [20], containing a range of different human motions. In the model there are no dependencies between different angles, i.e., C is diagonal.

There is a trade-off between tracking robustness and generality of the tracker. This temporal update model is very general, and allows for previously unobserved motion types, in contrast to action-specific motion models learned from training data, e.g. [23]. However, it does not guide the tracking to any large degree.

(a) $\mathbf{D}_t$ (b) $\hat{\mathbf{D}}(\boldsymbol{\alpha}_t)$ (c) $\mathbf{D}_t, \hat{\mathbf{D}}(\boldsymbol{\alpha}_t)$ difference

Figure 2. Image likelihood. (a) Observed human silhouette. (b) Silhouette rendered using the state $\boldsymbol{\alpha}_t$. (c) Silhouette difference: $\bullet = \neg(\mathbf{D}_t \cup \hat{\mathbf{D}}(\alpha_t))$, $\bullet = \mathbf{D}_t \cap \hat{\mathbf{D}}(\alpha_t)$, $\bullet = \mathbf{D}_t - (\mathbf{D}_t \cap \hat{\mathbf{D}}(\alpha_t))$, $\bullet = \hat{\mathbf{D}}(\alpha_t) - (\mathbf{D}_t \cap \hat{\mathbf{D}}(\alpha_t))$. *Best viewed in color.*

3.2. Image Likelihood Model

The image observation $\mathbf{D}_t$ consists of the human silhouette in a single image at time t. In the generative likelihood model $p(\mathbf{D}_t \,|\, \boldsymbol{\alpha}_t)$, this is compared to a silhouette $\hat{\mathbf{D}}(\boldsymbol{\alpha}_t)$ rendered by projecting all body parts onto the image plane (Figure 2). The likelihood of a certain $\mathbf{D}_t$ given $\boldsymbol{\alpha}_t$ is

$$w_{a,t} \propto e^{-\beta_a (1 - \frac{|\mathbf{D}_t \cap \hat{\mathbf{D}}(\alpha_t)|}{|\mathbf{D}_t \cup \hat{\mathbf{D}}(\alpha_t)|})} \quad (3)$$

where $\beta_1 > \beta_2 > \ldots > \beta_A$ are the smoothing parameters at different annealing levels a. During each timestep, the filter goes through all A levels in turn with Monte Carlo resampling between each level.

It should be noted that monocular tracking with an articulated 3D model and a general temporal update model like the one above, using silhouette information only, is a highly under-determined problem. In this paper, we do not study the performance of this tracker per se, but rather how the performance of a weakly constrained general tracker can be improved by including object contact information.

4. Human-Object Contact in Tracking

If the human holds an object in their hands, it will affect the image likelihood estimation since the human model can not explain all parts of the observed silhouette (Section 3.2). However, if the object instead is explicitly represented, it can be used to *help* the tracking. Object pose information can be employed in two different manners: 1) By explaining non-human foreground areas in the likelihood estimation; 2) By constraining the human body pose estimation during temporal update.

The underlying assumption here is that the grasped object is easier to track. This is true in general, since the object has fewer degrees of freedom (6 if rigid), and most often is easier to model visually thanks to straight edges, known texture etc. We do not further consider the object tracking problem here, concentrating instead on how the object pose can help the human tracking.

4.1. Human-Object Likelihood

As with the human tracking, (rigid) object tracking gives at each timestep t the transformation matrix T_t^{go} from the global coordinate system to the local object coordinate system. This makes it possible to determine the global position of object surface points as $\mathbf{t}_t^g = T_t^{go}\mathbf{t}^o$. The object surface can thus be projected onto the image plane, rendering a silhouette $\hat{\mathbf{D}}(\alpha_t, T_t^{go})$ jointly with the human model. This enhanced silhouette estimate can be employed in the human tracker likelihood model, Eq (3).

4.2. Human-Object Contact Constraints

Although it is possible to model contact with objects in the world at all points on the human, we currently consider hand-object contact (grasping) only. We furthermore limit ourselves to one elongated object. It is quite straightforward to extend the reasoning to two or more objects.

With one elongated object, the possible human-object contact states are 1) no object contact, 2) left hand-object contact, 3) right hand-object contact, 4) both hands-object contact. This can be encoded with two extra dimensions in the state $\boldsymbol{\alpha}_t$: left hand contact point λ_t and right hand contact point ρ_t, each with the values `NaN` in case of no contact. The parameter λ_{t-1} is propagated as:

$$\begin{aligned}
&\text{if } \lambda_{t-1} = \texttt{NaN} \\
&\quad \lambda_t = \begin{cases} z^o, & \text{hand} - \text{object dist} < T_E \\ \texttt{NaN}, & \text{otherwise} \end{cases} \\
&\text{else} \\
&\lambda_t = \begin{cases} \texttt{NaN}, & \text{w/ prob } p_{\text{letgo}} \\ \lambda_{t-1} + \nu\ ,\ \nu \sim N(0,\sigma), & \text{otherwise} \end{cases}
\end{aligned} \tag{4}$$

where z^o is the Z coordinate of the contact point in the object coordinate system, p_{letgo} is the probability of letting go of the stick, σ is a standard deviation describing how fast the hands can slide along the object, and T_E is a threshold. We discuss below how the hand-object distance is found. The parameter ρ_{t-1} is propagated similarly.

Assume now that the human is holding the object in the left hand. A point on the human's left hand is then in contact with a point on the surface of the object. If the object pose is known, this poses a kinematic constraint on the human model. The angles that fulfill the kinematic constraint lie on a 15D manifold in the 18D space spanned by the human torso and arm angles together with the position of contact on the object surface. This can be encoded in the temporal update model so that angles are only propagated on this manifold:

$$\boldsymbol{\alpha}_t = \mathcal{K}(\boldsymbol{\alpha}_{t-1}, T_t^{go}, \boldsymbol{\nu}) \tag{5}$$

where $\mathcal{K}$ is a non-linear function encoding the kinematic constraint, and $\boldsymbol{\nu}$ is a noise term.

We take a rejection-sampling-based approach to solving for the kinematic constraint. We define an error for the left side as

$$E_\lambda = \|T_t^{glh}\mathbf{t}^{lh} - T_t^{go}\mathbf{t}_t^o\| \tag{6}$$

where T_t^{glh} is the global to left hand transformation, $\mathbf{t}^{lh}$ the point of object contact in the left hand coordinate system, and $\mathbf{t}_t^o = [0, 0, z_t^o, 1]^T$ the point of left hand contact in the object coordinate system. For each particle $\tilde{\boldsymbol{\alpha}}_{t-1}^i$, all angles are propagated by random sampling according to Eq (2). Until the error $E_\lambda^i < T_E$, the sample $\boldsymbol{\alpha}_t^i$ is rejected and a new one drawn from the temporal model. This corresponds to rejecting those samples that move too far away from the kinematic constraint manifold in the state space. An analogous error, E_ρ, is defined for the right side. The kinematic constraints arising from right arm-object contact are accounted for in the same manner.

In addition to the kinematic constraints imposed by known grasping points, constraints are also introduced from the assumption that the object can not pass trough the human limbs, and vice versa. Object-body intersections are easily computed. The inter-penetration constraint is implemented in the same manner as the kinematic constraints; by rejecting and resampling the particles that do not obey it.

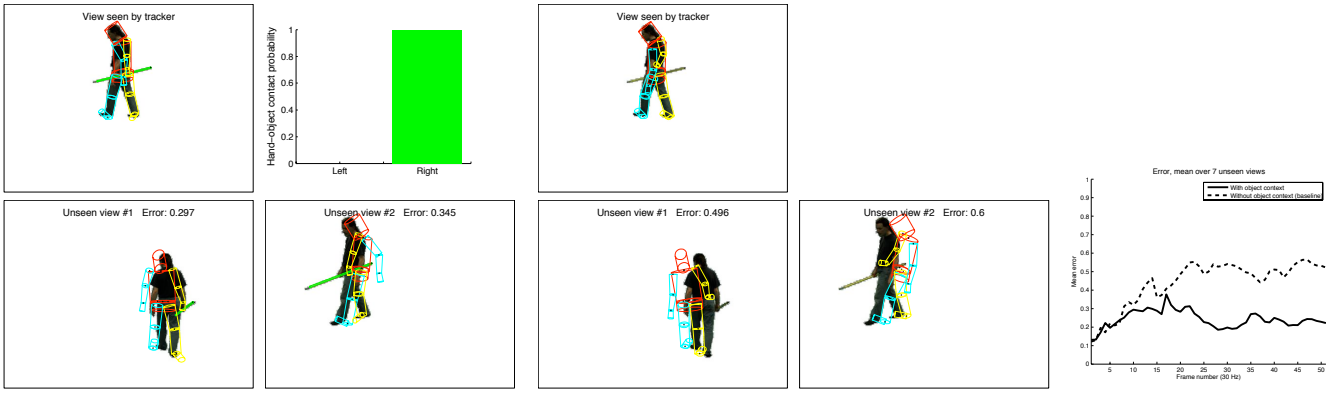
5. Experiments

The method is implemented in Matlab and evaluated on calibrated monocular image sequences of two different humans holding a stick, performing motions of varying complexity.[1] The performance of the human tracker, taking stick 3D pose into account, is evaluated against a baseline consisting of the same human tracker but with no object context. In all experiments the number of particles was $N = 100$. The human model is initialized with the true pose and true hand-object contact state in the first frame of each sequence.

Since the stick is thin in comparison to the limbs, the silhouette likelihood measure (Figure 2a) is not greatly affected by the stick. The effects on the likelihood of incorporating the object in the rendered silhouette (Figure 2b) are therefore not evaluated here. The likelihood is in all cases computed without taking the stick into consideration. Instead, the experiments evaluate how taking grasped objects into account increases the accuracy and robustness of tracking complex human motion.

In the experiments, the rigid object is also tracked using an APF as described in Section 3. The object (a stick) is modeled as one cylinder with 6 degrees of freedom. Since object tracking is not the focus of this paper we use all 8 camera views and 500 particles to obtain accurate object pose. The same silhouette likelihood is used as is used for

[1] The proportions of the human limbs in the cylindrical model vary slightly between different subjects.



(a) Frame 21 (b) Frame 21, baseline (c) Error over time

(d) Frame 41 (e) Frame 41, baseline (f) Hand-object contact over time

Figure 3. Global pose. Object context gives cues about the global position and orientation or the human. (a) Frame 21 with object context. (b) Frame 21 without object context (baseline). (c) Reconstruction error over time, mean over 7 unseen views; the view used for tracking is not included. (d) Frame 41 with object context. (e) Frame 41 without object context (baseline). (f) Left and right hand-object contact probability over time. (a,b,d,e) Color coding: ● = object, ● = head and torso, ● = left arm and leg, ● = right arm and leg. *Best viewed in color.*

the human body. As discussed in the Conclusions, we are currently developing a more sophisticated object tracking.

The probability of letting go of the stick with either hand is $p_{\text{letgo}} = 0.01$. According to Eq (4), this means that 1% of the particles in each resampling are released from the kinematic constraints for each hand. If those hypotheses correspond to reality, i.e., if the human has let go of the object with that hand, the hypothesized hand positions are free to move away from the object. If they move to the new actual hand position, and receive a higher likelihood weight, they will be more likely to survive resampling than particles that still encode hand-object contact. Similarly, according to Eq (4), particles that have hand positions closer to the object than $T_E = 100\text{mm}$ will be tied to the object. If this assumption is true, i.e., if the human has grasped the object, those hypothes will move with the object and receive a higher likelihood weight than particles that move randomly.

Tracking accuracy is evaluated in terms of negative log likelihood of the estimated pose $\boldsymbol{\alpha}_t$ in 7 image views *not used for tracking*. Thus, the reconstruction error for $\boldsymbol{\alpha}_t$ is

$$\varepsilon_t = \frac{1}{7}\sum_{u=1}^{7}\left(1 - \frac{|\mathbf{D}_t^u \cap \hat{\mathbf{D}}^u(\alpha_t)|}{|\mathbf{D}_t^u \cup \hat{\mathbf{D}}^u(\alpha_t)|}\right) \tag{7}$$

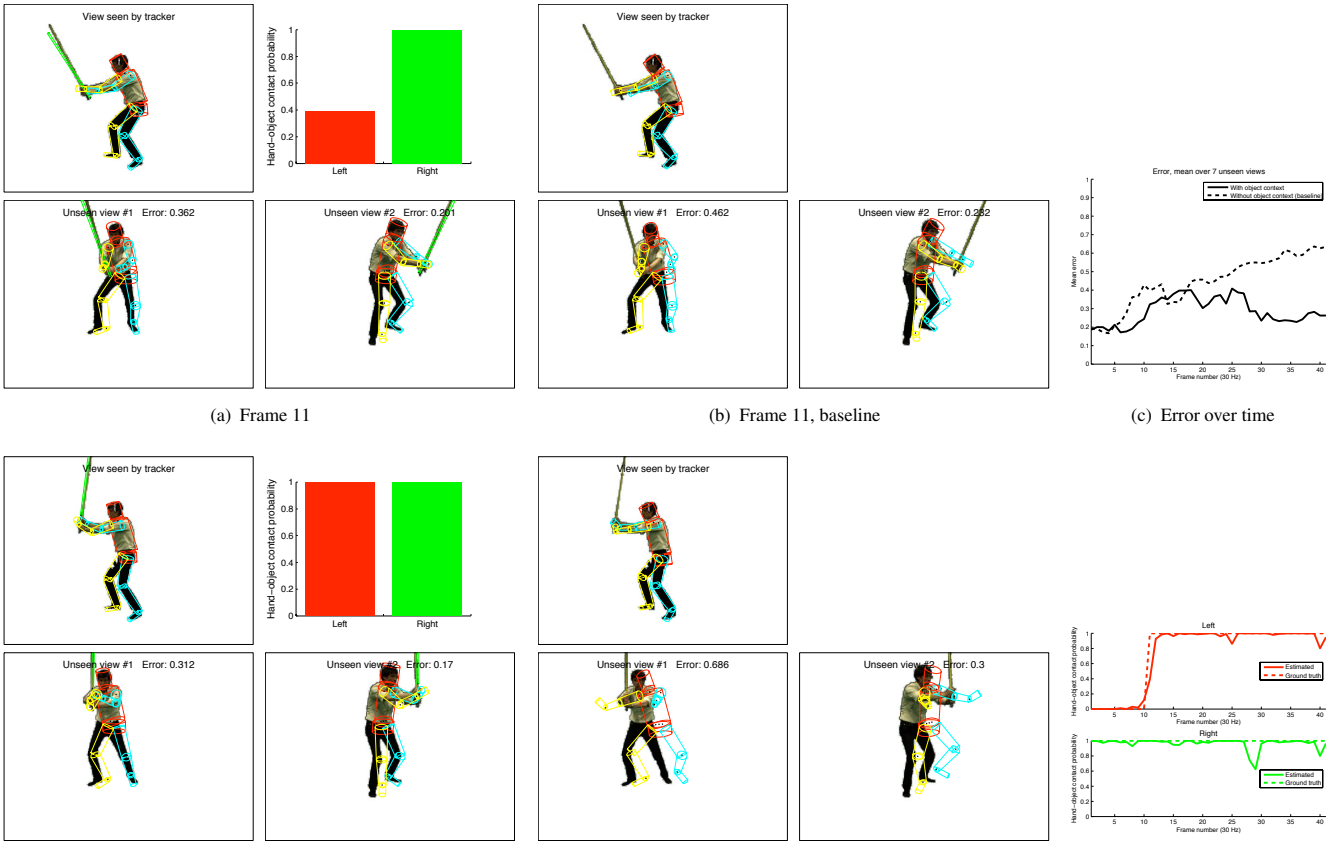
where $\mathbf{D}_t^u$ is the observed silhouette in unseen view u, and $\hat{\mathbf{D}}^u(\alpha_t)$ is the silhouette in view u generated from $\boldsymbol{\alpha}_t$.

Image data was captured using 8 Point Grey Grasshopper cameras (GRAS-20S4C), hardware synchronized at either 15 or 30 fps (4D View Solutions, France). One of the image views was used for monocular human tracking, while the others were used for evaluation of the tracking performance. Image resolution was 1624x1224 pixels. Calibration was performed using the Matlab Calibration Toolbox [4].

5.1. Global Pose

In this experiment, we use a sequence showing a human walking, carrying a stick in their right hand. Figure 3 shows the tracking result with and without object context.[2]

[2]For videos of all results see `www.csc.kth.se/~hedvig/cvpr10.html`



(a) Frame 11 (b) Frame 11, baseline (c) Error over time

(d) Frame 31 (e) Frame 31, baseline (f) Hand-object contact over time

Figure 4. Modeling grasping and letting go. The probabilities of left and right hand-object contact are automatically updated during tracking. (a) Frame 11 with object context. (b) Frame 11 without object context (baseline). (c) Reconstruction error over time, mean over 7 unseen views; the view used for tracking is not included. (d) Frame 31 with object context. (e) Frame 31 without object context (baseline). (f) Left and right hand-object contact probability over time. (a,b,d,e) Color coding: ● = object, ● = head and torso, ● = left arm and leg, ● = right arm and leg. *Best viewed in color.*

After 20 frames, the tracking error (Figure 3c), when object context is used, is about half of the error in the baseline case (no object context). Visual inspection of some of the unseen views (Figure 3a,b,d,e, lower views) shows the global position and vertical orientation if the torso is better estimated when object context is used; the 3D object pose helps disambiguate the human-camera distance estimate.

However, the estimate of leg pose relative to the torso is not greatly improved by object context; the left/right ambiguity in the silhouette (Figure 3d,e, upper views) causes a mix-up of the left and right legs in the 3D model, both with and without object context. One way to address this problem (apart from an action specific motion model, a more elaborate likelihood model, or a more detailed body model) is to include information of contact between the feet and the ground surface [26, 27].

The probability of hand-object contact is correctly estimated (Figure 3f), apart from a deviation in frame 47 due to failure in the object tracking.

5.2. Modeling Grasping and Letting Go

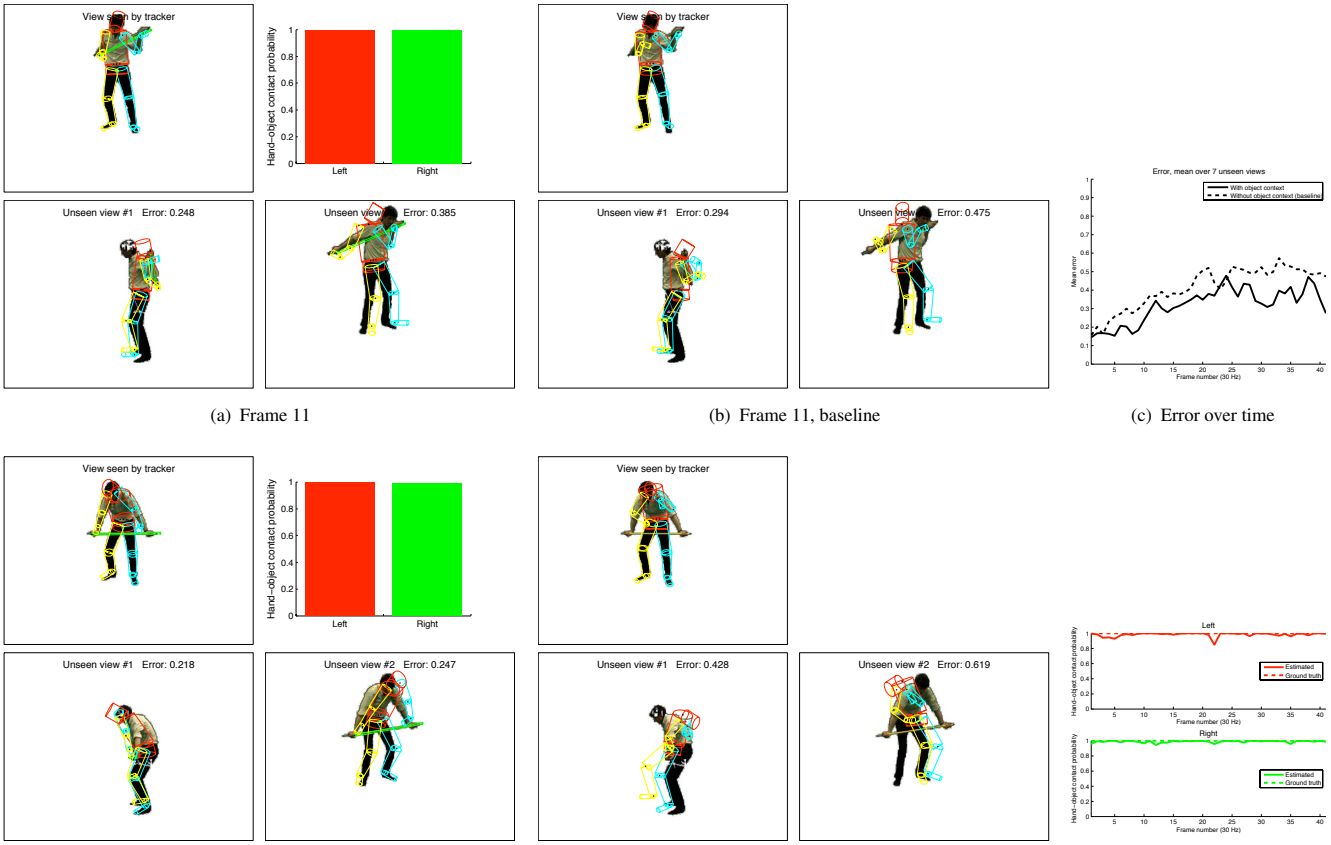
The sequence used in this experiment depicts a human holding a stick in their right hand (frame 1), grasping the stick with their left hand (frame 11), then moving the stick to the right (frame 31). Figure 4 shows the tracking result.

Firstly it can be noted that the hand-object contact is modeled correctly (Figure 4f), with a lag of approximately 1 frame when the left hand grasps the stick in frame 11.

Secondly, the hand-object contact constraint improves the accuracy of the arm pose estimation (Figure 4d,e), as can be expected. The reconstruction error (Figure 4c) is significantly lower when object context is used, partly due to the improved arm estimate, but also due to the improved estimate of global torso position and orientation (Figure 4d,e).

5.3. Complex Motion - "The Stick Trick"

To test the limits of the method, a sequence was captured of a human performing "the stick trick". This is a far more



(a) Frame 11 (b) Frame 11, baseline (c) Error over time

(d) Frame 41 (e) Frame 41, baseline (f) Hand-object contact over time

Figure 5. Complex motion - "The Stick Trick". The human is holding stick with both hands. (a) Frame 11 with object context. (b) Frame 11 without object context (baseline). (c) Reconstruction error over time, mean over 7 unseen views; the view used for tracking is not included. (d) Frame 41 with object context. (e) Frame 41 without object context (baseline). (f) Left and right hand-object contact probability over time. (a,b,d,e) Color coding: ● = object, ● = head and torso, ● = left arm and leg, ● = right arm and leg. *Best viewed in color.*

complex human motion than any of the examples found in the human tracking literature (Section 2). In fact, only a few humans can do the stick trick; it is a very awkward motion.

We use the first part of the sequence, showing a human holding on to the stick with both hands, palms facing forward, starting with the stick behind their back (frame 1), lifting the stick upward behind the back (frame 11), moving the stick forward above their head (frame 21), and lowering the stick in front of the human (frame 31).

In the end of this sequence, the upper arms are twisted approximately 120° inwards, and the wrists are twisted an additional 60° inwards, palms facing upward-outward. Thus, the human wrists and shoulders are twisted far beyond what is commonly assumed in articulated human tracking. Figure 5 shows the tracking result.

As with the two previous experiments, it can be noted that the hand-object contact probability is correctly estimated, and that object constraints help in estimating the global position and orientation of the arms and torso.

In this experiment with extreme arm motion, it is obvious (Figure 5b,e) that the baseline tracker without object contact constraints is unable to correctly reconstruct the arm pose. The object-constrained tracker is however able to maintain a qualitatively correct estimate of arm pose and twist relative to the torso (Figure 5a,d).

However, the reconstruction error is large both with and without object constraints, due to erroneous estimates of torso and head pose. The main reason is most probably the limitations of the human model, which has far fewer degrees of freedom than a real human body. This is particularly true for the torso; the very flexible human spine with its surrounding ribcage, flesh and muscles is modeled using a single truncated cone with three degrees of freedom relative to the pelvis. The model hands are also overly simplistic, modeled by one cylinder with one degree of freedom with respect to the lower arm. While the hands are not important in isolated full-body tracking, the hand model becomes central when grasping of objects is taken into account. An important topic for future work is thus to better model the human body, especially the torso and hands.

6. Conclusions

The key idea presented in this paper is that articulated 3D tracking of humans can be enhanced by taking into account knowledge about the pose of objects in the human's hands. It is noted that tracking of rigid objects is an easier task than articulated human tracking; objects in the human's hands can then be tracked independently of the human and then be used as "natural markers," giving cues about hand pose.

As described in Section 3, we use a standard formulation of human 3D tracking using an annealed particle filter with a linear temporal update model and a background difference likelihood. Inferred hand-object contact points impose kinematic constraints on the human model, which are encoded in the temporal update model using a rejection-sampling-based approach, described in Section 4.

Experiments in Section 5 show that the contact constraints improve single-view tracking performance for human motion, both in terms of accuracy (i.e., image distance from the true pose in other views) and robustness (i.e., the number of frames before tracking is lost). The kinematic constraints on the hand naturally help in the reconstruction of hand and arm pose, but also in reconstructing torso, pelvis and leg pose.

An important avenue for future work is to replace the cylindrical articulated human body model with a more flexible model with more of degrees of freedom. An example of such a model is described in [2]. Moreover, the impact of object constraints in different types of trackers and pose estimators will be studied. Future work will also explore contact relationships between body parts themselves (e.g. clasping of hands). Different types of grasps and the constraints they imply should also be modeled.

More elaborate object tracking methods are also being implemented, allowing for robust tracking of rigid objects with more complex shape and appearance.

Acknowledgments HK and DK were supported by EU IST-FP7-IP GRASP. MJB was supported in part by NSF IIS–0812364. We thank A. Weiss for help in acquiring the sequences used here and A. Balan for the base APF implementation.

References

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *CVPR*, 2004. 2

[2] A. Balan, M. J. Black, H. Haussecker, and L. Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *ICCV*, 2007. 2, 8

[3] J. Bandouch and M. Beetz. Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In *IEEE International Workshop on Human-Computer Interaction*, 2009. 2

[4] J.-Y. Bouguet. Camera calibration toolbox for Matlab. 5

[5] M. A. Brubaker and D. J. Fleet. The kneed walker for human pose tracking. In *CVPR*, 2008. 2

[6] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *ICCV*, 2009. 2

[7] D. Demirdjian, T. Ko, and T. Darrell. Constraining human body tracking. In *ICCV*, 2003. 2

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005. 2, 3

[9] Z. Duric, J. Fayman, and E. Rivlin. Function from motion. *PAMI*, 18(6):579–591, 1996. 2

[10] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, 2008. 2

[11] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. S. Davis. Context and observation driven latent variable model for human pose estimation. In *CVPR*, 2008. 2

[12] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 2

[13] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *ICCV*, 2009. 2

[14] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 2

[15] H. Kjellström, J. Romero, D. Martínez, and D. Kragic. Simultaneous visual recognition of manipulation actions and manipulated objects. In *ECCV*, 2008. 2

[16] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in computer vision-based human motion capture and analysis. *CVIU*, 104(2–3):90–126, 2006. 1

[17] R. Plankers and P. Fua. Visual interpretation of hand gestures for human-computer interaction: A review. *PAMI*, 25(9):1182–1187, 2003. 2

[18] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H.-P. Seidel. Markerless motion capture of man-machine interaction. In *CVPR*, 2008. 2

[19] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, 2000. 2

[20] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1–2):4–27, 2010. 3

[21] L. Sigal, S. Bathia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, 2004. 2

[22] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *CVPR*, 2001. 2

[23] R. Urtasun, D. J. Fleet, and P. Fua. Monocular 3D tracking of the golf swing. In *CVPR*, 2005. 2, 3

[24] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *CVPR*, 2006. 2

[25] M. Vondrak, L. Sigal, and O. Jenkins. The kneed walker for human pose tracking. In *CVPR*, 2008. 2

[26] M. Yamamoto and K. Yagishita. Scene constraints-aided tracking of human body. In *CVPR*, 2000. 2, 6

[27] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *PAMI*, 26(9):1208–1221, 2004. 2, 6