Probabilistic state space decomposition for human motion capture

Prabhu Kaliamoorthi, and Ramakrishna Kakarala

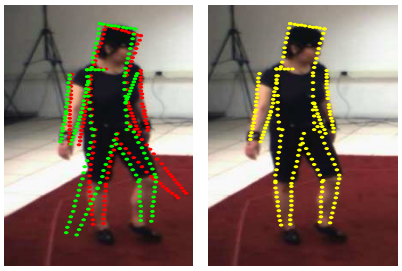
School of Computer Engineering, Nanyang Technological University, Singapore

Abstract. Model-based approaches to tracking of articulated objects, such as a human, have a high computational overhead due to the high dimensionality of the state space. In this paper, we present an approach to human motion capture (HMC) that mitigates the problem by performing a probabilistic decomposition of the state space. We achieve this by defining a conditional likelihood for each limb in the articulated human model as opposed to an overall likelihood. The conditional likelihoods are fused by making certain conditional independence assumptions inherent in the human body. Furthermore, we extend the popular stochastic search methods for HMC to make use of the decomposition. We demonstrate with Human Eva I and II datasets that our approach is capable of tracking more accurately than the state-of-the-art systems using only a small fraction of the computational resources.

1 Introduction

Model-based methods for human motion capture (HMC) [1–6] rely on particle based systems that either perform global optimization within a restricted search volume or use a sequential Monte Carlo (SMC) style tracker. These are designed to be general optimization and tracking methods which are applied to HMC. However, articulated objects such as a human body, have a number of conditional independence properties. For example, one could assume that the head and the leg poses are conditionally independent given the pose of the torso. Existing particle based systems for HMC do not make use of these properties, i.e., they are incapable of extracting a good leg pose from a sample which has poor overall likelihood due to head pose. This is observed in Figure 1a. Due to occlusion, these independence assumptions may not hold in a single view. However, most state of the art systems for HMC operate in a multi-view scenario, where these assumptions can be made to improve the tracking performance.

Partitioned sampling [7] is a technique that enables articulated object trackers to decompose the high dimensional state space. It has shown a 50% reduction in the tracking overhead for an articulated hand [7] with fewer (4) degrees of freedom than a human (25-50). However, annealing based methods such as Annealed Particle Filter (APF [1]), Interacting Simulating Annealing (ISA [3]) have been claimed to perform better than partitioned sampling for high dimensional spaces such as those used in HMC. Though recent studies [4] indicate partitioned sampling to be a promising alternative for HMC, state of the art systems



(a) (b)

Fig. 1: Part a) shows two poor overall poses which have a good fit for specific limbs, and b) shows a new pose extracted from the two poor poses that has an overall good fit.

[3, 4] do not make use of such a hierarchical decomposition, due to the lack of a systematic framework.

In this paper, we present a systematic approach to perform hierarchical decomposition. Though we apply our framework to HMC here, our method is a very general one, and it can be applied to other articulated objects such as a human hand or quadrupeds as well as other optimization problems that can be factorized. Our novel contributions are as follows. We describe a probabilistic framework to decompose the high dimensional state space of the HMC systems into subspaces of smaller dimension. The decomposition enables partitioned sampling type algorithms for HMC. We extend stochastic search methods (APF, ISA) to make use of the decomposition.

We validate the proposed method using the data from the Human Eva I and II datasets. Our results show that by decomposing the state space, we are able to capture complex human activity more accurately, using only a small fraction of the computational resources as the state-of-the-art systems [3, 6, 8].

2 Previous Work

There is a large body of research on HMC. Approaches such as [9, 10] use local optimization for tracking. Despite showing promising results, local optimization based techniques are known to get stuck in incorrect hypotheses [8]. Hence these methods are expected to fail when the model is not exact, or if the image features are uncertain. Even if these assumptions do hold, these methods could still benefit from a decomposed search framework such as the one we propose here. Discriminative methods [11, 12] that learn a mapping function between the image features and the human pose are known to require extensive training, and are expected to be sensitive to the appearance of the subject. Moreover, the generalization of these methods to novel poses not part of the training database is unclear.

More recently, HMC using pictorial structures [13] and belief propagation [14], which loosely assemble human parts to a plausible pose, have been proposed. However, these methods cannot ensure anatomically correct reconstruction of the human motion. Furthermore, they are confined to crude models of the subject and cannot be extended to more general surface meshes [2, 3]. Our method enforces certain conditional independence assumptions similar to [14]. However, our approach is very different from [14] since we enforce hard rather than soft constraints between limbs, i.e., distances between connected limbs cannot change in our method. Furthermore, since the conditional independencies are induced by the kinematic model, our method has a different set of conditional independencies than model free methods such as [14].

Deutscher et al. [1] formulate HMC as a global optimization problem, and use randomized search algorithms that locate the global optimum in a restricted volume of the state space. Furthermore, they extend [1] their work using inspiration from genetic algorithms and perform crossover when generating new samples. Our method could be considered as an extension of this, where we perform crossover based on the fit of the individual parts. Gall et al. [3] describe a multi pass solution that perform a crude tracking with global optimization, which is later refined by a smoothing filter and local optimization. Sidenbladh et al. [15] describe a complete generative framework to model based HMC using the sequential Monte Carlo tracker. Our approach is compatible with these methods which are referred to as generative methods or analysis by synthesis framework in the literature. However, none of these methods have proposed a systematic framework to decompose the state space, which is the main focus of this paper.

In [6], authors describe a framework for the HMC of multiple subjects in parallel. This can be considered as a specific instance of our method where the likelihood for the two subjects are assumed to be independent. We show that by exploiting the conditional independence structure, the HMC of a single subject can be made more efficient. As a result, the general framework we present here can be used for the HMC of multiple subjects.

3 State space decomposition

3.1 Overview

Recent model-based methods for HMC [1–6], approach the problem as a dynamic maximum likelihood estimation of a high dimensional state space $\mathcal{X}$. In this paper, we make certain conditional independence assumptions about the likelihood, i.e., we assume the likelihood could be factorized into subspaces $\mathcal{X}_i, i \in \{1, \ldots, L\}$ of much smaller dimensions. The factorization is achieved by defining conditional likelihoods for each part rather than an overall likelihood as commonly done in model-based methods [3, 4]. The conditional likelihoods are composed of a model to observation and an observation to model matching cost. However, since the conditional likelihoods are defined for each part, we decompose the observation probabilistically in order to define the conditional likelihoods. Using our conditional likelihoods, pose inference can be achieved using well known algorithms

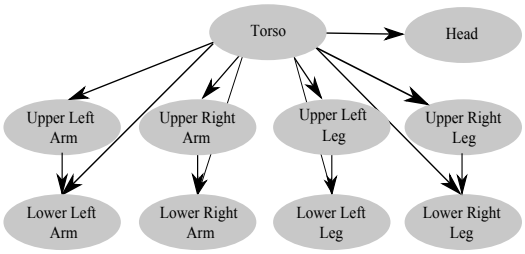


Fig. 2: The graphical model shows the conditional independence assumptions induced by the kinematic model.

such as non-parametric belief propagation (NBP) [14]. However, this would result in a Bayesian particle based system that approximates the posterior with a set of samples. Since most Bayesian methods such as [7, 15] require a very high number of samples for HMC, we extend the stochastic search methods for inference instead of Bayesian techniques. In the rest of this section, we describe our method in detail.

3.2 Marginal Likelihood

We achieve decomposition by making a number of conditional independence assumptions. The assumptions are induced by the kinematic model used for tracking. Figure 2 shows a directed graphical model representing the conditional independence assumptions made. The decomposed subspaces $\mathcal{X}_i$ are simply the degrees of freedom (DOF) for the rigid objects represented by the nodes in Figure 2. In order to illustrate model decomposition, let us consider the example of the lower left arm. After marginalizing the unrelated variables, the likelihood for the lower left arm and its parents is expressed as

$$P(lla, ula, tor) = P(lla|ula, tor)P(ula|tor)P(tor) \tag{1}$$

where lla, ula and tor represent the pose parameters corresponding to the lower left arm, upper left arm, and the torso respectively. In order to obtain the likelihood of lla one can marginalize the above equation as below

$$P(lla = l) = \int P(lla = l, ula = u, tor = t)\ du\ dt \tag{2}$$

The marginalization can be done numerically by a Monte Carlo approximation of the conditional likelihoods.

3.3 Conditional Likelihood

We use a variant of Oriented Chamfer Matching [16–18] in this work. However, the techniques that we discuss can be applied with other image features such as

(a) P(x|l=head, c) (b) Projection $\Pi(X)$ (c) Observation O

Fig. 3: The probability of the edge fragment given the part, $P(x|l,c)$, for head is shown in (a) (high probability in green and low probability in black). The edge fragments for the projection $\Pi(X)$ and the observation O, color coded according to the most probable label assignment $P(l|x,c)$ are shown in (b) and (c) respectively.

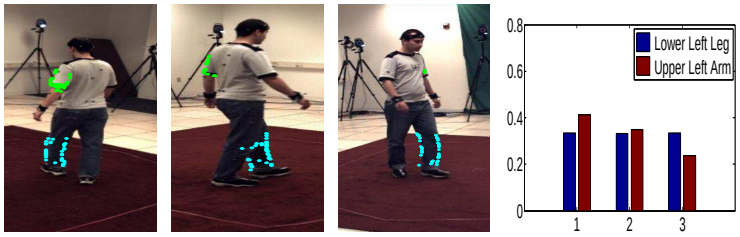
those used in [3, 4, 6]. Let $\mathcal{P}$ be the space $\mathbb{R}^2 \times [0, \pi]$ representing oriented edge fragments. Let $O^c = \{o_i \in \mathcal{P}\}$ and $\Pi^c(X) = \{\pi_i \in \mathcal{P}\}$ represent the respective sets of observation and synthesized edge fragments for a specific camera c. The oriented chamfer distance between the two for the camera c is defined as

$$\begin{aligned}\psi^c(O^c, X) = \frac{1}{|O^c|} \sum_{o_i \in O^c} d(o_i, \Pi^c(X)) \\ + \frac{1}{|\Pi^c(X)|} \sum_{\pi_i \in \Pi^c(X)} d(\pi_i, O^c)\end{aligned} \tag{3}$$

where $d : \mathcal{P} \times \mathcal{P}^N \rightarrow \mathbb{R}$, is a distance measure between an element in $\mathcal{P}$ and a set $\{\mathcal{P}\}$, $|O^c|$ and $|\Pi^c(X)|$ are the cardinalities of the sets O^c and $\Pi^c(X)$ respectively. It can be observed that ψ^c is composed of a model to observation and an observation to model matching term. The overall distance measure ψ is defined as the mean of the measure ψ^c from all cameras. The measure ψ provides a scalar cost that measures the overall match between the observation and the projection. In order to decompose it, we introduce a label l, which is distributed according to the L valued categorical distribution, formally, $l \sim Cat(L, p)$. The probability of the label indicates the degree of membership of an edge fragment to the individual rigid parts. Let us assume that the probability of the part given an edge fragment x in a camera c, $P(l|x,c)$, is known (described in section 3.4). Assuming an uniform prior over the edge fragments given the camera, and using Bayes rule, one can obtain $P(x|l,c)$.

$$P(x|l,c) = \frac{P(l|x,c)}{\sum_x P(l|x,c)} \tag{4}$$

Since the possible values x can take is dependent upon c, here we assume the



(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) $P(c|l)$

Fig. 4: The probability of the camera given the part, $P(c|l)$, for the lower left leg and the upper left arm is shown. The edge fragments (in color) which are most likely to be from the upper left arm and the lower left leg are shown in green and cyan respectively.

parameters for c and x are consistent, i.e., the probability $P(x, c)$ is non zero. Figure 3a shows the probability of the observed edge fragments for the head. In addition, assuming an uniform prior over the cameras and using Bayes rule, one can obtain the probability of the camera given the part as below

$$P(c|l) = \frac{\sum_x P(l|x, c)}{\sum_{x,c} P(l|x, c)} \tag{5}$$

$P(c|l)$ measures which camera is more likely to view a body part, and hence is more likely to help infer that part. Figure 4 shows $P(c|l)$ for the lower left leg and the upper left arm. It can be observed that for the lower left leg, which is equally visible in all three cameras, the measure $P(c|l)$ is equally distributed. Whereas, for the upper left arm which is nearly occluded in the third and second camera, the measure $P(c|l)$ is small for the third and second camera.

Using these probabilities the cost for a part l and camera c is expressed as

$$\phi_l^c(O^c, X) = E_{P(x=o_i|l,c)}[d(o_i, \Pi^c(X))] + E_{P(x=\pi_i|l,c)}[d(\pi_i, O^c)]] \tag{6}$$

where the first and the second term correspond to the observation to model and the model to observation matching cost respectively. Expressed differently, rather than summing the distance contribution from all the edge fragments as done in [16–18], we take a weighted sum with the weights estimated using Eq. 4. The total cost for the part l is expressed as the expectation of ϕ_l^c with respect to $P(c|l)$. Formally,

$$\phi_l(X) = E_{P(c|l)}[\phi_l^c(O^c, X)] \tag{7}$$

Modeling the likelihood to consider the visibility of the part is a novel aspect of our framework. Model-based methods in the current literature, take an average over all cameras to obtain the overall likelihood, which is equivalent to considering $P(c|l)$ to be uniform.

The conditional likelihood of a part given its parents is expressed as the cost for the respective part. Intuitively, treating the cost as conditional likelihood makes sense, since the cost for a part is conditioned on the value of the parameters for the parent links in the kinematic model. Formally,

$$-\ \log P(X_j|parents(X_j)) = \phi_l(X) + C \tag{8}$$

where X_j is a vector of parameters associated with the link j, $parents(X_j)$ is a vector of parameters for the ancestors of the link j (in the graphical model), and C is the normalization constant. For example, if j is considered to be the *lla* in Figure 2, then the vector X_j would be the parameter for the 1D joint associated with the *lla*, and the vector $parents(X_j)$ would comprise of the parameters for the 9 DOF for the *torso* (6) and the *ula* (3).

The formulation of the conditional likelihood in Eq. (8) is only an approximation due to occlusion. However, this is not a bad approximation since we take expectation over multiple cameras. Furthermore, since $P(c|l)$ is not uniform, the cost ϕ_l for the part l is more influenced by the view in which the part is not occluded.

3.4 Edge fragment prior

We estimate the prior probabilities for the observation edge fragments using the prior state estimate. A similar decomposition is performed in [6]. However, in our work the observation is a set of oriented edge fragments, whereas in [6] it is a silhouette. Let $\bar{X}$ be the prior estimate of the state. The edge fragments corresponding to the different parts can be separated by analyzing the part labels during the synthesis. Let $\Pi_l^c(\bar{X})$ represent a set of synthesized edge fragments corresponding to limb l and camera c. Let $\Pi^c(\bar{X})$ represent the complete set of synthesized edge fragments for camera c. The label probability for the observation edge fragment o_i is formally expressed as

$$\log P(l|o_i, c) = C\ -\frac{1}{T}\begin{cases} \sum\limits_{o\in O^c,\ p\in\Pi^c(\bar{X})} \frac{d_{\mathcal{P}}(o,p)}{|O^c||\Pi^c(\bar{X})|}, & \text{if } \Pi_l^c(\bar{X}) = \emptyset \\ d(o_i, \Pi_l^c(\bar{X})), & \text{otherwise} \end{cases} \tag{9}$$

where T is a constant used to control the uncertainty and C is the normalization constant. The set $\Pi_l^c(\bar{X})$ is empty when the part l is occluded in the camera c. For such an occluded part, we define the probability to be a low nonzero value. Since setting a constant value can make it sensitive to the distance measure being used, we define it to be the mean distance between the set O^c and $\Pi^c(\bar{X})$. The function $d_{\mathcal{P}}$ is a distance metric between oriented edge fragments [16], which is typically a convex combination of the Euclidean metric and orientation distance.

The constant T has a significant impact on the edge prior. As T approaches ∞, the edge fragments become equally likely to be from any part and as T approaches 0, the edge fragments are assigned to a single part with high probability. In general, we observe that reducing T improves the search performance. This is

expected since as T approaches 0, the edge prior is more informative. Therefore, the decomposed likelihood is influenced by both the model to observation and the observation to model matching cost. Whereas when T approaches ∞, as a result of a close to uniform edge prior, the observation to model matching cost in the decomposed likelihood is ineffective. However, when the observation is highly ambiguous, reducing T causes the tracker to get stuck in incorrect hypothesis. Hence T should be chosen as a trade-off between the two extremes.

The label probabilities for the projected model is defined as the Kronecker delta, since the part assignment is known with probability 1. Formally,

$$P(l = j | x = \pi_i, c) = \delta(\pi_i^l - j) \tag{10}$$

where π_i^l is the label corresponding to the projected model edge fragment π_i. The respective Figures 3b and 3c show the most probable label assignment for the synthetic output and the observation for a specific camera.

3.5 Inference

We adapted the stochastic search procedures (APF [1], ISA [3]) to make use of the decomposed likelihoods. In this section, we describe the modifications we made. Stochastic search methods start with the predicted estimate of the state for the current frame (obtained using GP regression [3] or simple motion prediction strategies such as constant velocity or position [1]) and construct a sequence of layers. Each layer consists of a set of samples, each of which is a tuple consisting of the pose vector X^j and its respective normalized weight w^j. The index $j \in \{1, \ldots, N\}$, where N is the number of samples used in a layer. Sample weights w^j are obtained by evaluating the annealed likelihoods and normalizing as below

$$\log q^j = -\beta^l \ \psi(X^j), \quad w^j = \frac{q^j}{\sum_{k=1}^{N} q^k} \tag{11}$$

where β^l is the inverse annealing temperature for the layer chosen dynamically for each layer [1] or by a predetermined schedule [3]. In each layer, samples are selected according to their normalized weights. The selected samples are then perturbed by an adaptive diffusion kernel, and re-weighted to result in a normalized set of samples. At the end of the last layer, the expected state of the sample set is considered to be the estimate of the state.

The principal change we made to the procedure is in Eq. (11). The sample weights in the modified procedure are obtained by evaluating the annealed marginal likelihoods rather than the overall likelihood. Formally,

$$\log q_i^j = -\beta^l \ \psi_i(X_i^j), \quad w_i^j = \frac{q_i^j}{\sum_{k=1}^{N} q_i^k} \tag{12}$$

where the subscript i indicates that the sample weights are for the decomposed subspace and ψ_i is the corresponding negative log marginal likelihood obtained by numerical marginalization in Eq. (2). Using the decomposed weights, new

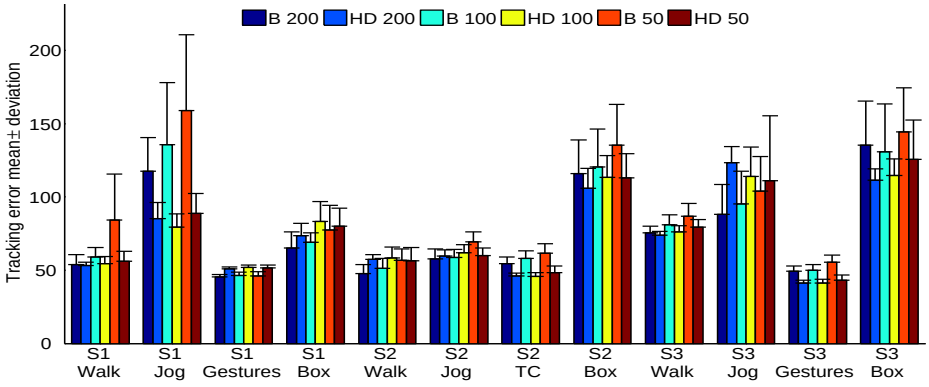


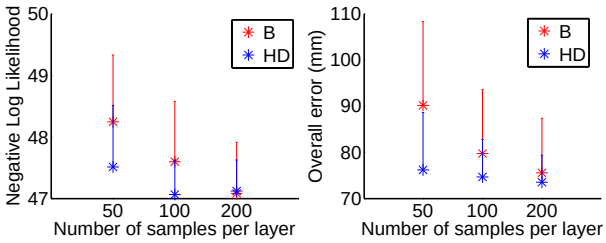
Fig. 5: Quantitative tracking results for Human Eva I dataset. It can be observed that our method (HD) performs competitively with 200 and 100 particles, and better than the baseline (B) with 50 particles.

samples are generated by re-sampling different X_i^js according to their respective w_i^js. They are finally combined to produce new samples X^j. This operation is similar to the crossover performed in [1], where the authors resample individual scalar values that makeup the state-space vector independently using the sample weights. However, we perform crossover based on the fit of the individual parts, i.e., cluster of scalar values (X_i^js) in the state space vector are resampled independently using different sample weights (w_i^j).

4 Experiments and Results

In this section, we first validate the proposed method by comparing it to a baseline configuration using the Human Eva I dataset. Then, we compare the proposed method with the state of the art methods used in HMC [3, 4, 6, 8] using the Human Eva II dataset. For Human Eva I, input from 3 RGB cameras and 2 grayscale cameras was used and for Human Eva II, input from all the 4 cameras was used for tracking. Similar to [8, 4], we registered a set of markers provided by the ground truth for the first frame to the model, in order to measure tracking error. The marker location in subsequent frames were used to measure the error. For Human Eva II, the online evaluation system [4] was used to estimate the tracking error. We used the parameters described in [18] for the distance mesaure on oriented edge fragments.

We used a kinematic model made of 10 rigid links (L) and 25 DOF for Human Eva I. We used an extended kinematic model, with ankles, for Human Eva II resulting in 12 rigid links and 27 DOF. We used ISA configured with the parameters used in [3] as the baseline procedure for our experiments in the Human Eva I dataset. We used a likelihood based on OCM [16, 17]. In our experiments, we found that for the torso, the cost ψ_{torso} is poorly constrained.



(a) S3 Jog NLL (b) Overall results

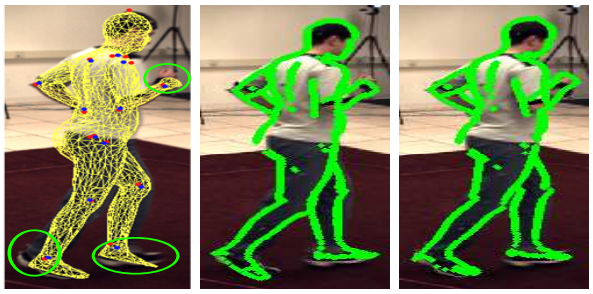
Fig. 6: The negative log likelihood (NLL) for the S3 Jogging sequence and the overall tracking results of all the sequences for the six configurations are shown in sub-figures (a) and (b) respectively. Mean error is shown by the asterisk, mean error plus one standard deviation is shown by the bar.

Hence we added the cost corresponding to the parts directly connected to the torso such as the head, upper arms and the legs to ψ_{torso}, in order perform a stable inference. Figure 5 shows the time averaged mean error and deviation ($+1\sigma$) of 5 different runs for 12 different sequences in the Human Eva I dataset. In the figure, the baseline algorithm is referred to as **B** and the decomposed search is referred to as **HD**. The number of samples used per layer is displayed next to the name of the configuration for both the algorithms.

It can be noticed that the decomposition significantly improves the performance for sequences such as S1 Jog, S2 TC, S2 Box, S3 Gestures and S3 Box. However, for S3 Jog, it can be observed that it performs worse. Furthermore, it can be observed that tracking error is worse when higher number of samples are used for tracking. On analysis, we found that when using decomposed search the model got stuck in incorrect hypotheses. However, the incorrect hypothesis had a higher likelihood. This is observed in Figure 6a, which shows the ensemble and time averaged negative log likelihood (NLL) for the S3 Jog sequence. It can be observed that the decomposed search still has a lower cost, i.e., it performs the task of search effectively. In general we observed that the decomposed search had a very wide effective search volume. Consequently, the decomposed tracker takes a different trajectory in comparison to the baseline version. We believe this results in marginally higher tracking error in sequences such as the S1 Gestures, S1 Box, and S2 Walk. However, these artifacts are caused by aspects such as poor model and observation, rather than the search procedure itself. This claim is further strengthened by the Human Eva II results that we present later which uses an accurate model and a relatively less noisy observation.

The overall mean error and deviation for all the 12 sequences is shown Figure 6b. It can be noticed from the average performance, that both the mean error and the deviation are reduced in comparison to the baseline method. In addition, it can be noticed that the performance of the decomposed search is not

significantly affected by the reduction in number of samples as opposed to the baseline method.



(a) [3] (b) HD SA (c) HD S

Fig. 7: Qualitative comparison of the tracked output for the subject S4, frame 580, camera 1. It can be observed that the leg pose estimated by our method is better than those reported in [3].

We compare our results with [3] for Human Eva II, since studies such as [6, 8] use the implementation of [3]. Furthermore, the study in [4] uses an approximate model and hence it results in much higher tracking error than [3]. Rather than implementing the method in [3] ourselves, we compare our approach to the L1 results reported in [3], which is equivalent to our search procedure. We used the surface mesh provided with the Human Eva II dataset for tracking. Similar to [3], we reduced the mesh to have 4000 triangles to have an acceptable computational load.

A high speed approximate method using KD trees was used to synthesize oriented edge fragments from the surface mesh. Figure 3b shows the synthesized edge fragments, which can be observed to contain a few incorrect occluded edge fragments. We obtained the probability in Eq. (10) from the skinning parameters [10] of the vertex. The distance $d_{\mathcal{P}}$ was robustified with the Geman-McClure function in order make the objective robust to outliers. Details of the implementation such as the parameters used can be obtained from the source code supplied with the paper.

The decomposed search procedure was configured to use 8 layers and 70 samples per layer. The annealing schedule parameter α for the decomposed search was set to 0.2 and the adaptive diffusion parameter γ was set to 0.4 [3]. The parameter T described in Section 3.4 was set to 15. We used a constant position model [1] for prediction and a simple silhouette extraction method. The observation set O^c included oriented edge fragments from the silhouette and gray image in the foreground region. Table 1 summarizes the tracking results for the two sequences compared, where our method is referred to as **HD SA** (since it uses silhouette and appearance). It can be observed that our procedure performs better than [3] on most slots.

If the appearance of the subject is highly textured, the noise can be significantly higher than methods such as oriented chamfer matching can handle. In such a scenario the silhouette alone is the most reliable image feature. Hence we ran the above procedure with the observation set O^c containing oriented edge fragments from the silhouette alone. The results for this test are summarized in Table 1 as **HD S** (since it uses only silhouette). It can be observed that even without using appearance related features, our procedure results in comparable tracking performance as [3]. We provide the baseline results using ISA and OCM with the parameters used in [3], as **B SA** in the table. It can be observed that the proposed method **HD SA** is significantly better than the baseline.

Figures 7 shows the tracked result superimposed on the observation from the S4 sequence. It can be observed that the leg pose estimated by our method is slightly better than that in [3]. Figure 8 shows the tracked result for the S2 and S4 sequences from the Human Eva II dataset. Videos of the tracked and smoothed results, as well as the source code used to generate them, are available online [19].

		S2			S4		
Frames		1-350	1-700	1-1202	2-350	2-700	2-1258
Absolute	[3]	41.5 ± 8.0	45.0 ± 12.9	43.8 ± 10.7	34.6 ± **4.6**	38.5 ± 6.9	38.1 ± 5.8
$\mu \pm \sigma$	HD SA	**37.0** ± 6.9	**41.7** ± **9.1**	**42.4** ± 9.6	**31.2** ± 5.4	**34.8** ± 6.4	**36.3** ± **5.7**
(mm)	HD S	39.4 ± 7.6	44.6 ± 12.6	46.6 ± 12.9	31.6 ± 5.1	35.3 ± **6.1**	37.1 ± 6.2
	B SA	39.2 ± **6.8**	44.7 ± 9.8	44.2 ± **9.2**	32.5 ± 5.9	36.7 ± 7.5	40.7 ± 9.9
Relative	[3]	45.8 ± **9.0**	48.4 ± 13.7	46.6 ± 11.4	43.9 ± 8.2	47.0 ± 10.6	45.3 ± 9.1
$\mu \pm \sigma$	HD SA	**41.4** ± **9.0**	**43.4** ± **8.9**	**45.2** ± **10.1**	**32.0** ± 5.9	**36.2** ± 7.6	**38.2** ± **7.4**
(mm)	HD S	44.8 ± 10.3	47.9 ± 13.1	50.2 ± 14.3	32.4 ± **5.7**	37.0 ± **7.3**	39.0 ± 7.9
	B SA	45.0 ± 9.9	48.6 ± 10.7	48.4 ± 10.2	32.7 ± 6.5	38.3 ± 9	43.4 ± 13.5

Table 1: Our tracking results for the Human Eva II dataset are presented next to those reported in [3] for the subjects S2 and S4. The absolute and the relative error were obtained using the online evaluation system [4]. Best result in **bold**.

Method	Samples	Computation time per frame
[3, 6, 8]	3750	76 sec
ours	560	6 sec

Table 2: Number of samples used and the computation time on a standard PC. It can be observed that our method uses less than one-sixth of the number of samples used in [3].

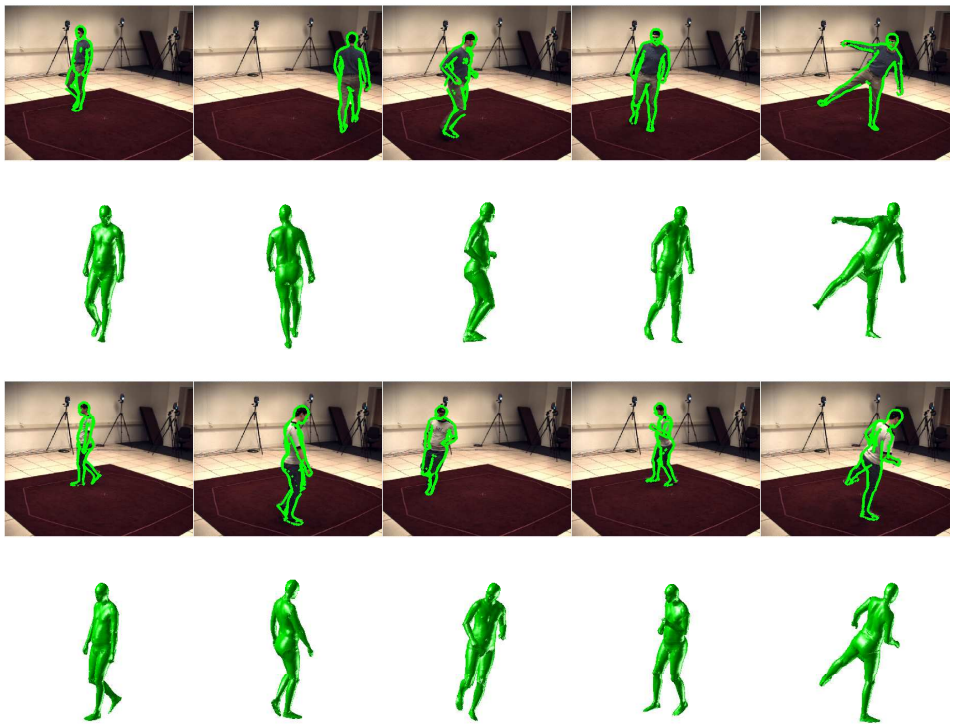


Fig. 8: Tracking results for S2 and S4 sequences from Human Eva II. Odd rows show the model output superimposed on the input from the camera, even rows show the 3d model.

The decomposition procedure we introduce in this paper marginally adds to the per sample overhead, but we found that this is insignificant in comparison to the rest of the processing. Since computational overhead to HMC is directly related to the number of samples used [4], it can be significantly reduced by using our method. The number of samples used for tracking and the computation time on a standard PC for the Human Eva II dataset is shown in Table 2. It can be observed from the table that the proposed method uses less than a sixth of the samples used in [3]. The computation time of our method is significantly lower both due to the decomposed search method which requires significantly lower number of samples and the OCM based likelihood, which can be realized at high speeds. Furthermore, the implementation in [3, 6, 8] uses a GPU based rendering. We did not use GPU acceleration in any form. We believe that by using a GPU based implementation of our method, HMC with the accuracy achieved in [3, 6, 8] would be possible at few frames per second.

5 Conclusion and Future Work

In this paper, we describe a probabilistic framework to decompose the high dimensional state space of the human motion capture system. We show that by defining conditional likelihood for each limb rather than an overall likelihood for the human model, a number of conditional independence assumptions can be made that enable the decomposition of the state space. We extend the state-of-the-art search method for HMC to make use of the decomposed subspaces. We demonstrate using the Human Eva I and II datasets that the decomposition framework significantly improves the tracking performance per sample, enabling the search technique to reach the tracking performance reported in the state of the art systems using only a fraction of the computational resources. In this work, we apply the decomposed search for the HMC of a single subject. In the future, we hope to apply our framework to multiple interacting subjects such as in [6], and for articulated hand tracking.

References

1. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. IJCV **61** (2005) 185–205
2. Sigal, L., Balan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS. (2007) 1337–1344
3. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. IJCV **87** (2010) 75–92
4. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV **87** (2010) 4–27
5. Pons-Moll, G., Baak, A., Gall, J., Leal-Taixé, L., Müller, M., Seidel, H.P., Rosenhahn, B.: Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling. In: ICCV. (2011) 1243–1250
6. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: CVPR. (2011) 1249–1256
7. Maccormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: ECCV. (2000) 3–19
8. 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. (2009) 1746–1753
9. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: CVPR. (1998) 8–15
10. Ballan, L., Cortelazzo, G.M.: Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In: 3DPVT. (2008)
11. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In: CVPR. (2004) 882–888
12. Bo, L., Sminchisescu, C.: Twin Gaussian processes for structured prediction. IJCV **87** (2010) 28–52
13. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR. (2010) 623–630

14. Sigal, L., Isard, M., Haussecker, H.W., Black, M.J.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. IJCV **98** (2012) 15–48
15. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: ECCV. (2000) 702–718
16. Liu, M.Y., Tuzel, O., Veeraraghavan, A., Chellappa, R.: Fast directional chamfer matching. In: CVPR. (2010) 1696–1703
17. Shotton, J., Blake, A., Cipolla, R.: Multiscale categorical object recognition using contour fragments. TPAMI **30** (2008) 1270–1281
18. Kaliamoorthi, P., Kakarala, R.: Directional chamfer matching in 2.5 dimensions. IEEE Signal Processing Letters **20** (2013) 1151–1154
19. (https://sites.google.com/site/prabhukaliamoorthi/publications) [Online; accessed 6-September-2014].