Combining Simple Models to Approximate Complex Dynamics

Leonid Taycher, John W. Fisher III, and Trevor Darrell
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology Cambridge, MA, 02139
{lodrion,fisher,trevor}@ai.mit.edu

No Institute Given

Abstract

Stochastic tracking of structured models in monolithic state spaces often requires modeling complex distributions that are difficult to represent with either parametric or sample-based approaches. We show that if redundant representations are available, the individual state estimates may be improved by combining simpler dynamical systems, each of which captures some aspect of the complex behavior. For example, human body parts may be robustly tracked individually, but the resulting pose combinations may not satisfy articulation constraints. Conversely, the results produced by full-body trackers satisfy such constraints, but such trackers are usually fragile due to the presence of clutter. We combine constituent dynamical systems in a manner similar to a Product of HMMs model. Hidden variables are introducied to represent system appearance. While the resulting model contains loops, making the inference hard in general, we present an approximate non-loopy filtering algorithm based on sequential application of Belief Propagation to acyclic subgraphs of the model.

1 Introduction

It has long been conjectured that human success in interpreting complex dynamic scenes (e.g. nonrigid objects undergoing complicated motion) is due to our ability to simultaneously perform inference at multiple logical scales or levels of abstraction [12, 10]. On the other hand, most model-based machine vision algorithms – both deterministic and stochastic – approach this task with monolithic top-down frameworks using a single state parameterization.

Many of these algorithms have been based on filtering in hidden Markov chain frameworks and share the “generate-and-test” method of observation likelihood computation based on probabilistic generative models. In such systems, a (usually deterministic) low-level process is used to extract visual image features. The model state is estimated using a “generate-and-test” framework, where the features corresponding to a particular state are produced by a deterministic function and compared to features extracted from the image, and the result is then converted into likelihood based on the sensor noise model. These approaches are popular in part because inference in Markov chain models has been extensively studied, and exact inference algorithms are available once the model has been defined. On the other hand, approximations are involved in *defining* the model are severe (approximate dynamics and observation likelihoods are usually used). These methods are biologically implausible, since they use a “feed-forward” information flow from low-level to high-level abstraction levels, while it has been shown that the high-level feedback is important for human perception [10].

In this work, we propose an alternative approach to tracking complex systems using multiple representation framework. Instead of committing to a single representation with complicated dynamics, our framework combines multiple stochastic trackers, each using simple approximate dynamics and estimating the state of the body at a particular representation (e.g. levels of the hierarchical structure or levels of abstraction of [13]). The consistency of estimates is enforced by communication between individual trackers through shared appearance representations. Rather than competing to provide the best explanation for the observations (i.e. Switching Models paradigm [17]), our simple models cooperate in the spirit of Product of Experts [8] and Product of HMMs [3] frameworks to maximize the probability of the observations (Figure 1(c)). The major difference between our approach and HMM variants, such as Factorial HMMs [7] and Coupled HMMs [1] is that rather than partitioning the state vector and semi-independently evolving its parts, we maintain an overcomplete state representation. The power of the framework arises from the interactions between individual Markov chains that redundantly describe the system.

For example, the human body exhibits behaviors that are subject to constraints at different scales, and dynamics described in a single parameterization may be quite complex (e.g. a straight-line motion of the hand involves complicated interactions between joint angles [16]). Intuitively, explicitly modeling dynamics at each scale is more appropriate than using the common technique in which a single level of abstraction (and its unique parameterization, usually joint angles) is selected, and inference is performed only at this level (Figure 1(a)). Such algorithms utilize the hierarchical structure of the observed body only to simplify rendering of appearance features corresponding to particular states, which are then used for state likelihood computations.

Our approach results in a model that has complicated structure but simple potentials, compared to standard models with simple structure but complicated potentials. While exact inference in the resulting loopy model is complicated, we show that an approximate on-line filtering algorithm based on Belief Propagation [18] can successfully infer system state.

2 Prior Work

Dynamical system state estimation in complex models is an important problem in machine vision (e.g. object pose estimation and tracking). It is commonly posed as a search problem in a high-dimensional configuration space, and has been studied extensively in the context of the human body tracking algorithms. While deterministic optimizations (e.g. gradient descent) have been investigated [5, 2, 19], the need to model uncertainty (especially important for the monocular case with depth ambiguities) resulted in the advent of stochastic algorithms.

Most of these algorithms have been based on filtering in hidden Markov chain frameworks and share the "generate-and-test" method of observation likelihood computation. The visual features corresponding to a particular state are produced by a deterministic function and compared to features extracted from the image, and the result is then converted into likelihood based on the sensor noise model. Various appearance features, and 3D body descriptions have been proposed, cf. surveys in [6, 15, 14, 21].

Multiple functional forms of the approximations to state posterior and state dynamics likelihood distributions have been used. Early human tracking approaches [20] used a Kalman Filtering framework, thus implicitly modeling all constituent distributions with Gaussians. These assumptions have been later relaxed to account for nonlinear monomodal dynamics (while still using Gaussian state model) by using Extended [9] or Unscented [25] Kalman Filters. Switched linear systems [17] were proposed to describe arbitrary learned dynamics.

Different approaches have been used for modeling dynamics in systems with sample-based state distribution representations. The original CONDENSATION algorithm and its variants were used in [22, 11]. Importance sampling with proposal distribution defined over hand motion manifold was proposed in [29]. The hybrid Monte Carlo sampling that used observed image to modify sample locations (not only sample weights) was presented in [4]. Partitioned [11] and Layered [26] sampling used factored state dynamics, which allowed semi-independent propagation of parts of the state vector. Several techniques [24, 23] have been proposed for modifying state dynamics (or rather diffusion) parameters based on the previous state distribution. An implicit dynamics model presented by [21] uses a motion-capture database to predict the future state distribution.

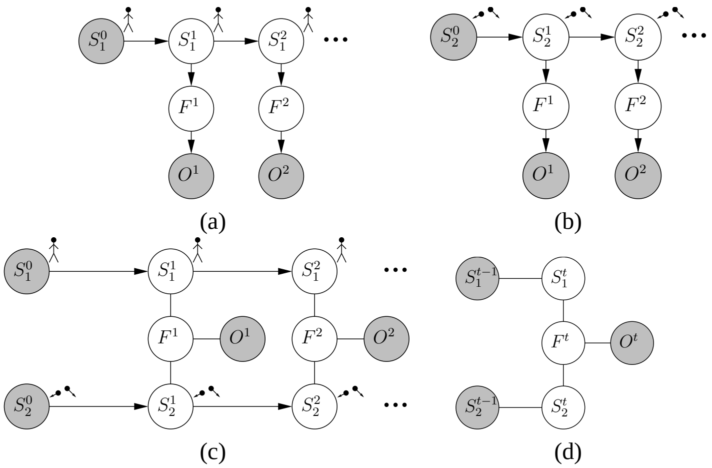
Our approach is to combine weak cooperating trackers. In this sense it is similar to Bayesian modality fusion [28], which combines the output of head trackers operating in different modalities based on individual trackers' reliability.

We propose to combine weak trackers using the product model proposed in [8] and [3]. The Product of Experts paradigm [8] uses "simple" (expert) probability distributions to model more complex ones by taking their products and renormalizing rather than summing the constituent distributions (the mixture model). The product model accepts (assigns high probability) only those areas of the parameter space that all experts accept.

Our model uses renormalized products of tractable probability distributions that are derived from individual trackers, each modeling a low-dimensional constraint on the data. (note that we use PoHMM [3] conventions for combining directed and undirected links in the Graphical Model (Figure 1(c))). In this work, we assume that constituent stochastic trackers are completely specified, and the appearance feature hierarchy (if any) is known; we concern ourselves with inference on the combined model, rather than learning its structure.

3 Probabilistic Formulation

A common approach to state estimation in complex systems is the generate-and-test framework. In this framework, state likelihood is computed as the similarity between appearance features (e.g. color blobs, edge pixels, volumetric models) extracted from



(a) (b) (c) (d)

Fig. 1. Combining weak trackers. (a) An articulated body tracker with state parameterized by the joint angles, and with known dynamics $p(S_1^t|S_1^{t-1})$, feature-generation model $p(F|S_1)$ and feature-observation model $p(O|F)$. If inference is performed on the individual chain, then appearance features may incorporated into the observation likelihood as $p(O|S) = \int_F p(O|F)p(F|S)$. (b) Independent object tracking systems with dynamics $p(S_2^t|S_2^{t-1})$ and feature-generation model $p(F|S_2)$ (see text for details); the feature-observation model is shared with the first system. (c) Combined model with potentials $\phi(S_1^t, S_1^{t-1}) = p(S_1^t|S_1^{t-1})$, $\phi(S_2^t, S_2^{t-1}) = p(S_2^t|S_2^{t-1})$, $\phi(F, S_1) = p(F|S_1)$, $\phi(F, S_2) = p(F|S_2)$, $\phi(O, F) = p(O|F)$. The edges leading to F nodes lack edges to preserve independence relationships. (d). A tree-shaped subgraph on which approximate inference is performed. The marginal distributions, $p(S_1^{t-1}|O^{0..t-1})$ and $p(S_2^{t-1}|O^{0..t-1})$, have been computed at the previous iteration, and are not modified; O^t is observed.

the image and features generated for a particular value of the state. When the appearance generating functions are complex such systems become quite sensitive to observation noise, since small variations in the appearance features extracted from an observation may cause large changes in the state estimate. Thus the quality of state estimates depends on the quality of appearance estimation. Likelihood computation in the generate-and-test framework may be interpreted as implicit instantiation and marginalization over features (F) in the model shown in Figure 1(a) by specifying $p(O|S) = \int_F p(O|F)p(F|S)$.

By making appearance features an explicit variable in the model, we are able to provide an alternative interpretation of this framework as estimating the *intermediate features*, rather than the state, at each timestep. This insight enables us then to combine heterogenous models at the level of features and improve tracking.

The task of estimating appearance features at every timestep may be solved by using models with different state representations. For example, if we are interested in tracking flesh blobs corresponding to hands, we may use the full-body representation parameterized by joint angles or we may track hands as two independent objects moving with first-order dynamics. Other models are also possible. None of the models is exact (each uses approximated dynamics and appearance generation), and each has its own advantages and disadvantages. While tracking with the full-body model is complicated, its estimates of the hand blob positions would always satisfy constraints implied by the physics of the human body. On the other hand, independent blob trackers are usually robust, but may produce state estimates incompatible with the human body. Different models have different failure modes, so combining them may improve an overall estimate of the system appearance. In the case of stochastic trackers, such an appearance-estimation model can be represented with graphical models in Figure 1(a, b). For this example, F may be interpreted as flesh-blob positions explaining part of the observed image, S_1 as a set of joint-angle parameters in the full-body model, and S_2 as positions and velocities in the independently-moving-objects model); superscripts denote time.

In a single chain the probability distribution of latent appearance features $p(F|S,O)$ can be expressed as

$$p(F|S,O) = \frac{1}{Z} p(F|S) p(F|O) \tag{1}$$

and can be interpreted as a product of two "experts" [8], one predicting appearance based purely on the state, and the other based on the observations. We combine multiple chain models by using experts for all available models to predict appearance, in this example:

$$p(F|S,O) = \frac{1}{Z} p(F|S_1) p(F|S_2) p(F|O) \tag{2}$$

intuitively, the probability is high for a particular value of F only if *all* experts assign high probability to this value.

The resulting system is shown in Figure 1(c) (note that the edges leading to F^i are left undirected in order to preserve independence relationships). In contrast to conventional Product of HMMs [3], constituent chains are combined at the latent rather than observed variable. Since, as has been noted, quality of individual state (S_i) estimations depends on the appearance estimation, the combined system should improve both appearance and individual state estimates.

Yet another reason for combining individual models is that the exact dynamics and appearance generation distributions in these models are either unknown, or are so complicated that their exact modeling is currently infeasible. Thus, certain approximations have to be made for tractability purposes (e.g. representing dynamics with simple parametric distributions and the appearance with simple volumetric models). These approximations increase the uncertainty of both likelihood (by increasing the number of modes [23]) and dynamics (by increasing the size of the parameter volume that has to be searched). Furthermore, these approximations may violate the independence relationships in the single-chain models (i.e. F^1 and F^2 are not in general independent conditioned on S_1^1 and S_1^2). A multi-chain combined model captures dependencies between F nodes more accurately.

Although we discuss the case when individual models predict the same features, it is possible to combine models with intersecting feature sets. In that case the combined feature model would be the union of individual feature sets, and the likelihood potentials are extended to produce uniform likelihoods for features that are not part of an original submodel. If there exists a hierarchy of appearance descriptions (e.g. joint angles, 3D volumetric descriptions of individual segments, and textured region representations of the human body), then systems that approximately describe behaviors at each level in the hierarchy may also be combined.

3.1 Inference in the Multi-chain Model

Single-chain models are popular because there exist efficient inference algorithms for them. While our proposed multi-chain model is loopy (Figure 1(c)) and exact inference is notoriously complicated in such models, we take advantage of the fact that we are interested only in marginal distributions for the state nodes to propose an efficient algorithm for *filtering* in multi-chain model.

Let us consider the model in Figure 1(c). At time $t = 1$, we can ignore the nodes with superscripts (times) $t \geq 2$. If the initial states S_1^0 and S_2^0 are independent (as shown), then the resulting subgraph is a tree, and so we can use standard Belief Propagation technique to compute the marginal distribution at state nodes S_1^1 and S_2^1.

$$p(S_1^1|O^1) = \frac{1}{Z} \int_{S_1^0} \phi(S_1^1, S_1^0) p(S_1^0) \int_{F^1} \left[\phi(F^1) \phi(F^1, S_1^1) \int_{S_2^1} \left[\phi(F^1, S_2^1) \int_{S_2^0} \phi(S_2^1, S_2^0) p(S_2^0) \right] \right], \tag{3}$$

where $\phi(F^1) = \phi(O^1, F^1)$ (the equivalent expression of $p(S_2^1|O_1)$ is not shown).

Filtering at the next timestep ($t = 2$) is more complex since the model now contains loops and the exact inference would require representing the joint $p(S_1^1, S_2^1|O^1)$:

$$\begin{aligned} p(S_1^2|O^1, O^2) = \frac{1}{Z} \int_{F^2} \Bigg[\phi(F^2) \phi(F^2, S_1^2) \int_{S_2^2} \Bigg[\phi(F^2, S_2^2) \\ \int_{S_1^1, S_2^1} \phi(S_1^2, S_1^1) \phi(S_2^2, S_2^1) p(S_1^1, S_2^1|O^1) \Bigg] \Bigg] \end{aligned} \tag{4}$$

In order to simplify computations, we would like to approximate the joint distribution $p(S_1^1, S_2^1|O^1)$ with a product $q(S_1^1)q(S_2^1)$. It is easily shown that the best such approximation (in the KL-divergence sense) is the product of marginal distributions $p(S_1^1|O^1)$ and $p(S_2^1|O^1)$. Substituting $p(S_1^1|O^1)p(S_2^1|O^1)$ for $p(S_1^1, S_2^1|O^1)$ in Equation 4, we obtain an approximate inference equation:

$$\begin{aligned} p(S_1^2|O^2) \approx \frac{1}{Z} \int_{S_1^1} \phi(S_1^2, S_1^1) p(S_1^1|O^1) \int_{F^2} \Bigg[\phi(F^2) \phi(F^2, S_1^2) \\ \int_{S_2^2} \Bigg[\phi(F^2, S_2^2) \int_{S_2^1} \phi(S_2^2, S_2^1) p(S_2^1|O^1) \Bigg] \Bigg] \end{aligned} \tag{5}$$

The similarity between Equations (3) and (5) suggests an approximate filtering algorithm that estimates marginal distributions of the state variables by recursively applying Belief Propagation to acyclic subgraphs of the form shown in Figure 1(d), using the marginal state distribution obtained at time $t - 1$ as priors at time t.

While the exact nature of the approximation involved requires further analysis, it may be shown that it preserves the main property of the exact model, that the appearance features that are assigned zero probability under one of the constituent models are not used in computation of either of the marginal distributions.

At each timestep approximate inference may be performed using message passing in 4 steps.

1. Compute $\mu_{S_1^{t-1} \to S_1^t} = \int_{S_1^1} \phi(S_1^t, S_1^{t-1}) p(S_1^{t-1} | O^{0..t-1})$ and $\mu_{S_2^{t-1} \to S_2^t} = \int_{S_2^1} \phi(S_2^t, S_2^{t-1}) p(S_2^{t-1} | O^{0..t-1})$,
2. Compute $\mu_{S_1^t \to F^t} = \int_{S_1^t} \phi(F^t, S_1^t) \mu_{S_1^{t-1} \to S_1^t}$ and $\mu_{S_2^t \to F^t} = \int_{S_2^t} \phi(F^t, S_2^t) \mu_{S_2^{t-1} \to S_2^t}$,
3. Compute $\mu_{F^t \to S_1^t} = \int_{F^t} \mu_{S_2^t \to F^t} \phi(O^t, F^t)$ and $\mu_{F^t \to S_2^t} = \int_{F^t} \mu_{S_1^t \to F^t} \phi(O^t, F^t)$
4. Compute marginal state distributions $p(S_1^t | O^{0..t}) \propto \mu_{S_1^{t-1} \to S_1^t} \mu_{F^t \to S_1^t}$ and $p(S_2^t | O^{0..t}) \propto \mu_{S_2^{t-1} \to S_2^t} \mu_{F^t \to S_2^t}$

If inference on constituent Markov chains was performed individually, it would still involve steps analogous to 1 and 4 (and partially 3), so combining models does not introduce much additional complexity to the inference process.

4 Implementation

In this work, we are interested in tracking human upper-body motion observed under orthogonal projection. To this end we define an upper-body articulated model with 13 degrees of freedom (2 in-plane translational dofs, 3 rotational dofs at the neck, 3 rotational dofs at each shoulder and 1 rotational dof at each elbow).

Since no parametric form is known for body-pose distribution, we choose to use a sample-based density representation. Common sample-based particle-filtering approaches (e.g. CONDENSATION), compute posterior state distribution at each timestep by sampling from distribution at the previous timestep propagated by dynamics, and reweight samples by their likelihood. If the configuration space is complex, then, unless the dynamics are well known, this procedure results in many samples falling into areas of zero likelihood, and thus increasing the number of samples that need to be drawn. An alternative is *likelihood sampling*[27], when pose samples are drawn from the pose likelihood function and are then reweighted based on the propagated prior. Although this method results in greater per-sample complexity, it enables us to use fewer samples, since they are in general placed more appropriately with respect to the posterior distribution.

To implement likelihood sampling, we take an advantage of the fact that we are able to not only evaluate, but also sample from observation likelihood definitions for the head and hands (in this case mixtures of Gaussians corresponding to face detector outputs and to detected flesh-colored blobs). We define observation likelihood using latent image observation likelihoods: face detector output for the head segment, flesh-color likelihoods for the hands, and occlusion edge map matching for the rest of the segments. Once the 2D face and hand position samples has been drawn, we use them together with inverse kinematics constraints to define pose-proposal distribution. We then use this distribution in the importance sampling framework to obtain sample from the pose likelihood.

We define our proposal distribution as in [27]. In defining the proposal distribution we take an advantage of the fact that once head and hand positions and neck configuration are specified, then arm configurations (shoulder and elbow angles) are independent, and each arm has only two degrees of freedom. The complete description of likelihood pose-sampling may be found in [27].

While a tracker based on the likelihood sampling is can successfully operate with a small number of samples and is self recovering, it is extremely sensitive to feature detector failures (such as flesh-color misdetections). In this work, we combine a likelihood-sampling tracker with low-level flesh-blob tracking using robust Kalman filtering. These tracking systems share appearance features (flesh-colored blobs), enabling us to combine them in the manner described in Section 3.

We have applied our dual-chain tracker to three sample sequences, with results shown in Figure 2. For each frame in the sequence we have rendered fourty randomly drawn samples from the posterior state distribution (the frontal view overlayed on top of the input image is shown in the middle row, and side view is shown in the bottom row). The tracking results for the first sequence are also available in the submitted video file (rendered at one third of the framerate). In most frames, the tracker succeeded in estimating poses (that contained significant out of plane components and self occlusions), and was able to recover from mistracks (e.g. around frame 61 in the third sequence).

While the current implementation of the tracker is not realtime (on the average 1-3 seconds per frame, mostly due to inefficient image computations), its performance should be compared to other tracking methods using the same dynamics and image likelihood models. In Figure 3, we compare the performance of the enhanced dual-chain tracker using 1000 samples per frame (first column), likelihood-sampling tracker using 1000 samples (second column), CONDENSATION tracker with 5000 samples that runs as fast as the dual-chain tracker (third column), and finally CONDENSATION tracker with 15000 samples (the smallest number of samples that enables CONDENSATION to perform with accuracy approaching dual-chain tracker performance). The results are presented using the same method as in Figure 2, the frontal view is shown overlayed on top of the input image, side view is shown to the right.

The dual-chain tracker was able to successfully track the body through the entire sequence. The likelihood-sampling tracker was generally able to correctly estimate the pose distribution, but failed on frames where image features were not correctly extracted (cf. frames 20, 22, etc). The CONDENSATION variant with 5000 samples failed after 30 frames due partly to sample impovereshment (note that only few distinct samples were drawn in frames 40 and later). Increasing the size of sample set to 15000 (with similar increase of the running time) allowed CONDENSATION to successfully track through more of the sequence.

5 Conclusions and Discussion

We have proposed a methodology for combining simple dynamical models with redundant representations as a way of modeling more complex dynamical structures such as a moving human body. The approach was motivated by the simple observation that nearly all “generate-and-test” approaches to tracking complex structures implicitly marginalize over an intermediate feature representation between state and observation. By making the feature representation explicit in our approach we obtained a straightforward means of mediating between simpler models as a means of capturing more complex behavior.

Exact inference on the resulting structure is complicated due to the introduction of loops in the graphical structure representing the combined models. However, as a consequence of the fact that we are primarily interested in the filtering (or tracking) problem, rather than the smoothing problem, an approximate inference method, based on sequential inference on acyclic subgraphs provides a suitable alternative to exact inference. This approximation has the important property that infeasible configurations in *any* of the naive models precluded an infeasible configuration in *all* of the others.

Empirical results demonstrated the utility of the method for tracking the upper body of a human. The method compares favorably with the well-known CONDENSATION algorithm in two ways. First, a monolithic approach using CONDENSATION required a significantly greater number of samples in order to explore the configuration space sufficiently as compared to the multi-chain method. Secondly, and perhaps more importantly, in the experiments presented the estimate of the posterior state distribution more accurately represents the uncertainty of the upper-body pose than the alternative methods. This is particularly encouraging considering the simplicity of combining constituent models as compared to a monolithic approach.

References

1. M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of IEEE CVPR'97*, 1997.
2. Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *Proc. of CVPR*, 1998.
3. Andrew Brown and Geoffrey E. Hinton. Products of hidden markov models. In *Proceedings of Artificial Intelligence and Statistics*, pages 3–11, 2001.
4. Kiam Choo and David J. Fleet. People tracking using hybrid monte carlo filtering. In *Proc. ICCV*, 2001.
5. D. Demirdjian. Enforcing constraints for human body tracking. In *Proceedings of Workshop on Multi-Object Tracking, Madison, Wisconsin, USA*, 2003.
6. Dariu Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
7. Zhoubin Ghahramani and Michael Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–273, Nov/Dec 1997.
8. Geoffrey E. Hinton. Products of experts. In *Proc, of the Ninth International Conference on Artificial Neural Networks*, pages 1 – 6, 1999.
9. Nebojsa Jojic, Matthew Turk, and Thoman S. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *Proc of International Conference on Computer Vision*, 1999.
10. Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 20(7), June 2003.
11. John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV (2)*, pages 3–19, 2000.
12. D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three dimensional structure. In *Proceedings of the Royal Society of London*, 1978.
13. David Marr. *Vision*. W.H Freeman and Company, New York, 1982.

14. Thomas B. Moeslund. The analysis-by-synthesis approach in human motion capture: A review.
15. Thomas B. Moeslund. Summaries of 107 computer vision-based human motion capture papers. Technical report, 1999.
16. P. Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42:223–227, 1981.
17. Vladimir Pavlović, James M. Rehg, Tat-Jen Cham, and Kevin P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *Proc. ICCV*, 1999.
18. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1997.
19. James M. Rehg and Takeo Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Fifth International Conference on Computer Vision*, pages 612–617, 1995.
20. K. Rohr. Towards models-based recognition of human movements in image sequences. *CVGIP*, 59(1):94–115, Jan 1994.
21. Hedvig Sidenbladh. *Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences*. PhD thesis, Royal Institute of Technology, Stockholm, 2001.
22. Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proc. European Conference on Computer Vision*, 2000.
23. Christian Sminchiesescu and Bill Triggs. Covariance scaled sampling for monocular 3d human tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.
24. Christian Sminchiesescu and Bill Triggs. Kinematic jump processes for monocular 3d human tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.
25. B. Stenger, P. R. S. Mendonca, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. *Proc. British Machine Vision Conference*, 2001.
26. J. Sullivan, Andrew Blake, Michael Isard, and John MacCormick. Object localization by bayesian correlation. In *ICCV (2)*, pages 1068–1075, 1999.
27. Leonid Taycher and Trevor Darrell. Bayesian articulated tracking using single frame pose sampling. In *Proc. 3rd Int'l Workshop on Statistical and Computational Theories of Vision*, Oct 2003.
28. Kentaro Toyama and Eric Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *ACCV'00*, 2000.
29. Ying Wu, John Y. Lin, and Thomas S. Huang. Capturing natural hand articulation. In *Proc. International Conference on Computer Vision*, 2001.

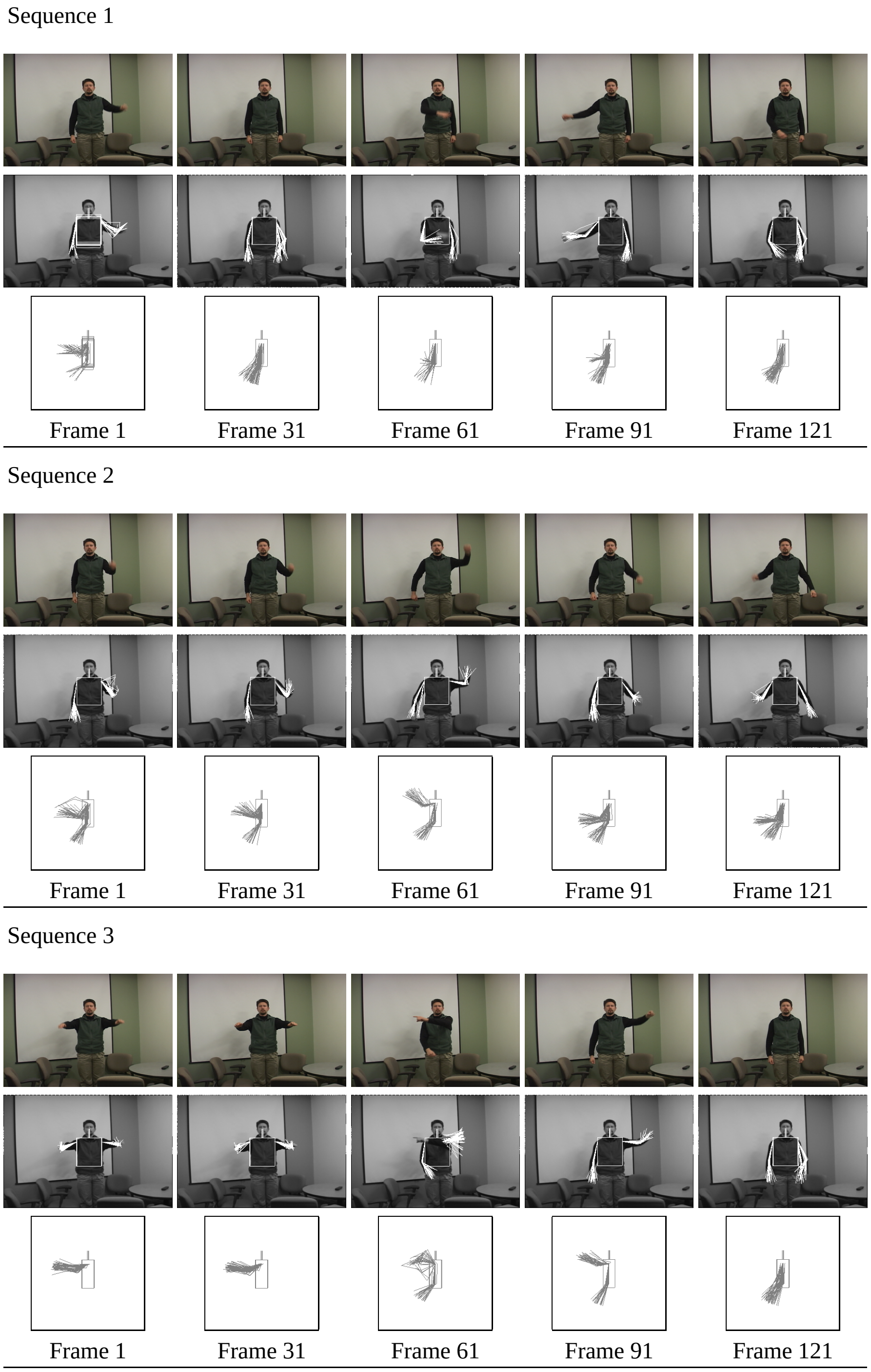


Fig. 2. Applying dual-chain tracking to three sample sequences. Five frames from each sequence are presented. The top row contains input frames. Fourty random particles from the estimated posterior pose distributions are shown: in the middle row, the patricles are rendered onto the input image (frontal view), and in the bottom row they are rendered in the side view. Note that while a mistrack has occured on the third sequence near frame 61, the tracker was able to recover.

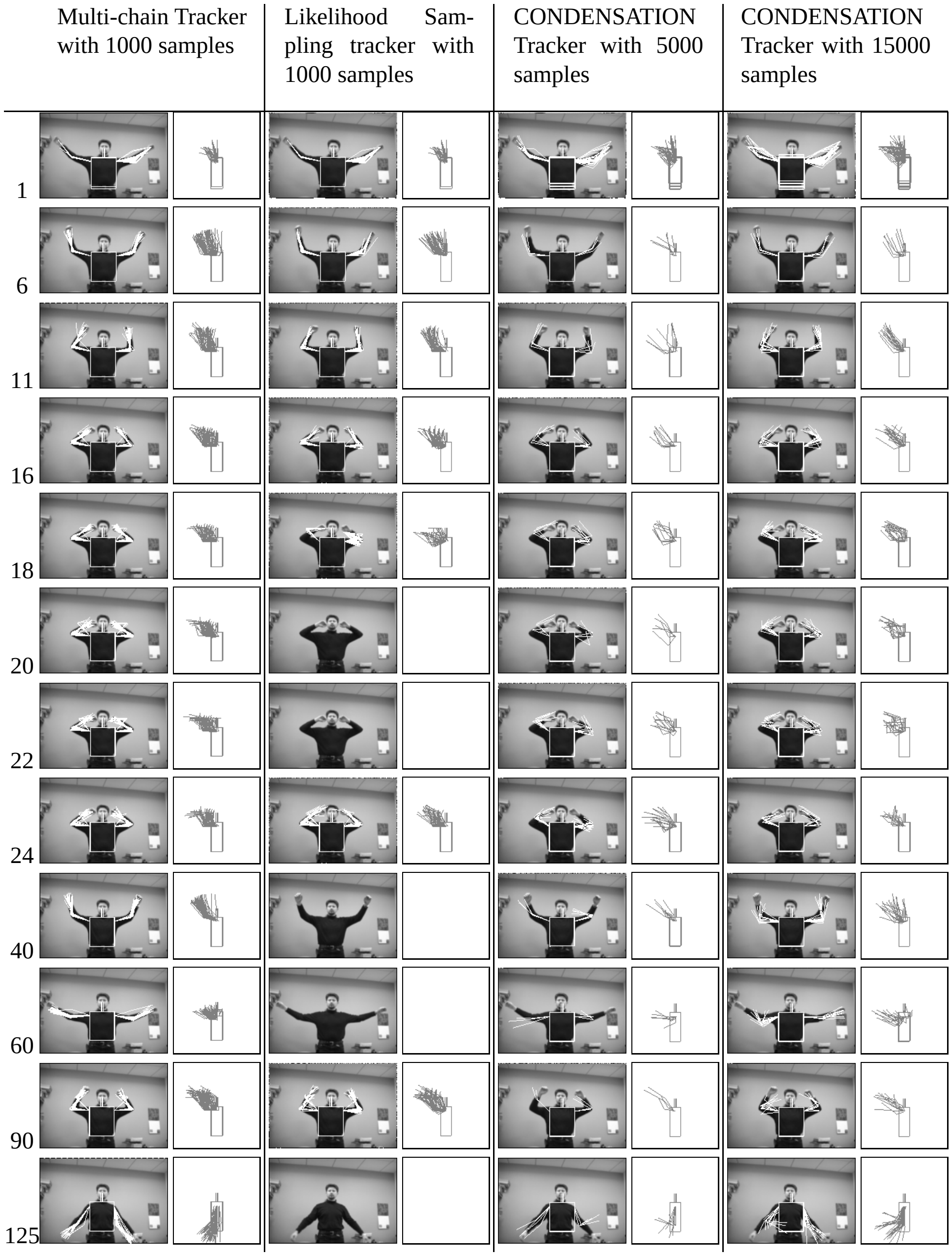

Multi-chain Tracker with 1000 samples
Likelihood Sampling tracker with 1000 samples
CONDENSATION Tracker with 5000 samples
CONDENSATION Tracker with 15000 samples
1
6
11
16
18
20
22
24
40
60
90
125

Fig. 3. Applying four tracking algorithms to a sample sequence. For each frame a set of fourty random pose samples were drawn from estimater posterior distribution and the corresponding skeletons was rendered (frontal view overlayed on the frame and side view next to it). Errors in feature detection caused likelihood-sampling tracker to fail on some of the frames (no samples were produced).