Articulated Particle Filter for Hand Tracking

German Ros*, Jesús Martínez del Rincón† and Ginés García Mateos‡
*Computer Vision Center and Computer Science Dpt., UAB, Spain
†Digital Imaging Research Centre, Kingston University, London, UK
‡Dpt. Informática y Sistemas, University of Murcia, Spain
E-mail: gros@cvc.uab.es

Abstract

This paper proposes a new version of Particle Filter, called Articulated Particle Filter —ArPF—, which has been specifically designed for an efficient sampling of hierarchical spaces, generated by articulated objects. Our approach decomposes the articulated motion into layers for efficiency purposes, making use of a careful modeling of the diffusion noise along with its propagation through the articulations. This produces an increase of accuracy and prevent for divergences. The algorithm is tested on hand tracking due to its complex hierarchical articulated nature. With this purpose, a new dataset generation tool for quantitative evaluation is also presented in this paper.

1. Introduction

Tracking articulated motion from video sequences is one of the most active areas in computer vision. It can be defined as the ability to recover a sequence of articulated poses, in most cases a human, from a video or sequence of images. This is critical in applications such as human computer interaction, motion capture, medical analysis and biometrics, to name a few.

Probabilistic tracking, such as the particle filter (PF) method, has been a promising approach for markeless articulated tracking [15, 3]. In initial approaches, the tracking complexity increased exponentially with the number of moving targets or independent segments. However, in recent years several methods have been proposed to alleviate this challenge. In this line, motion models [1, 11, 14] have been applied in order to provide a context which strong constrains. The disadvantage is a loss of generality, reducing the application to a pre-trained set of actions. More shopisticated models of interaction were proposed in [7] in order to account for complex interaction between parts.

Among the techniques that do not limit the generality, limb hierarchy [5] has been proposed to reduce the complexity of high-dimensional search. Since the search space is partitioned according to the limb hierarchy, each estimate constrains the possible configurations of subsequent segments in the articulated chain [5, 9]. However, an inherent problem of this approach is the need to estimate accurately the parameters of the initial segment to avoid the propagation of the error to the subsequent segments [6, 4]. In a similar fashion, partitioned sampling (PS) [8] reduces the complexity of searching in many dimensions by decomposing the space into sub-spaces which are estimated independently of one another. This is done without altering the probability density function, but the algorithm can only be applied when specific conditions hold, such as observation functions can be measured independently and the partition is meaningful [3, 8]. To overcame previous limitations, annealed particle filter (APF) [2] uses simulated annealing to guide hypotheses towards the global optimum, reducing the risk of getting stuck in local optima. APF can also be understood as a soft partitioning [3] of the search space as opposed to PS.

Despite the excellent results of APF in articulated motion [13], it is not appropriate for all kind of hierarchical spaces. The sampling of the search space when the segments in the articulated chain have different grade of variability is inefficient and leads to bad estimations. This is due to a bad estimation of the diffusion noise among layers, only based on the variance of the sampling set. In fact, we defend that such a variance is caused by two different factors, the spreading of the hypotheses and the reliability of the observation function. Therefore, in order to estimate a valid configuration, the algorithm requires to remove one of them from the estimation problem, by means of an unambiguous observation function [13, 11] or a articulated chain with equally variable segment [3]. Proof of this is the necessity of a multi view system [13, 11] and the inability

of APF to solve monocular sequences [13] for complex articulated chains.

In this paper, we propose a new method to track articulated models based on semi-hard partitioning of the search space. This partitioning is done according to the hierarchical properties of kinematic chains, and allows us to cope with imperfect observation functions. Experiments are performed on hand tracking using Kinect as a sensor to show its superior performance against APF in complex articulated scenarios.

2. Articulated Particle Filter

The Articulated Particle Filter (ArPF) is based on the well-known Annealed Particle Filter (APF), but it has been designed to cope with the limitations of APF for tracking hierarchical articulated chains.

APF follows the classical procedure defined by PF, but introduces the concept of multilayered search space in order to improve the convergence of particles toward good solutions. In this space the different layers represent smoothed versions of the original, guiding the propagation process from the smoothest layer (M) to the original one, through a set of levels. In this way, the method can obtain a "general view" of the search space, and therefore can avoid falling in local maxima. The set of layers is generated by modifying the behavior of the observation function $\omega(\mathbf{Z}, \mathbf{X})$, such that for the m-th layer $\omega_m(\mathbf{Z}, \mathbf{X}) = \omega(\mathbf{Z}, \mathbf{X})^{\beta_m}$, given that

$$\beta_0 > \cdots > \beta_M \tag{1}$$

where $\mathbf{X}$ denotes the model's configuration vector and $\mathbf{Z}$ notates the history of observations.

This process gives rise to a multi-layer sampling-propagation scheme as showed in Eq. (2) and (3).

$$\pi_{k,m}^{(i)} \propto w_m(\mathbf{Z}, \mathbf{X} = s_t^{(i)}), \sum_{i=0}^{N-1} \pi_{k,m}^{(i)} = 1 \tag{2}$$

$$s_{k,m}^{(i)} = s_{k,m-1}^{(i)} + \mathbf{B}_m, i = 0, \ldots, N-1 \tag{3}$$

where $s^{(i)}$ and $\pi^{(i)}$ are the state vector and the weight of each particle i at instant k, N is the total number of particles and m is the layer of the annealing process.

The term $\mathbf{B}_m$ models the diffusion noise of the particle set for the m-th layer, an is drawn from a normal distribution $\mathcal{N}(0, \mathbf{P}_m)$, where $\mathbf{P}_m$ is proportional to the covariance of the particle set. Setting the dispersion levels in this way produces a sort of automatic partition of the search space (soft partitioning), which allows for optimising subsets of parameters according to the variance measured in the particle set. Such an idea was presented in [3], claiming that the variance of any parameter is directly related to the degree of optimal estimation of such parameter. So, parameters with higher variance are considered as very influent and are firstly adjusted. In theory the idea can produce a better convergence of the algorithm.

However, the previous statement does not always hold, since the variance of a parameter is determined by both the accuracy of the hypothesis distribution and the discriminatory capacity of $\omega(\mathbf{Z}, \mathbf{X})$. In other words, in real conditions the observation function cannot produce correct measurements along the full range of a parameter, which implies that in some areas there is not enough information to estimate those parameters with no uncertainty. This phenomenon introduces a factor inversely proportional to the discriminatory capacity of $\omega(\mathbf{Z}, \mathbf{X})$ into the variance of those parameters. As a consequence high variances can mean that a given parameter is dominating the space (and more particles can be used to reduce its variance), or else, that the parameter is reaching its confusion lower bound (and no further improvement can be obtained). As conclusion, in order to efficiently move particles in the space, the diffusion noise of a parameter should be guided by a ratio between the variance of the whole set of particles S and the discriminatory capacity of the observation function.

Nevertheless, the discriminatory capacity of an observation function $\omega(\mathbf{Z}, \mathbf{X})$ is hard to define as it changes over time and depends on many uncontrolled variables, e.g. $\mathbf{Z}$ and $\mathbf{X}$. However a simplified version of this function can be defined as the resolution of the k-th parameter of $\mathbf{X}$ in the range $[a, b]$, given a specific case $\mathbf{Z}$. Here, D is the dimension of $\mathbf{X}$ extended with two extra dimension due to the interval parameters $[a, b]$ and $\delta_{\omega,Z}^k(\mathbf{X}, a, b)$ is the resolution of $\omega(\mathbf{Z}, \mathbf{X}^*)$ when all the parameters except the k-th are fixed (Eq. (4)).

$$\delta_{\omega,Z}^k(\mathbf{X}, a, b) = |b-a|\Big(1 - \big(\operatorname*{argmax}_{a \le i \le b}(\xi(k,i)) - \operatorname*{argmin}_{a \le j \le b}(\xi(k,j))\big)\Big) : \mathcal{R}^{D+2} \to \mathcal{R} \tag{4}$$

where $\xi(k, x) = \omega(\mathbf{Z}, \mathbf{X}^*(k) \leftarrow x))$ represents the assignment of optimal values for all the parameters X^* except for the free parameter k, which is adjusted by assigning it values in the range $a \le x \le b$. μ_k^* is the mean of $\delta_{\omega,\mathbf{Z}}^k(\mathbf{X}, a, b)$. In simple words, the resolution $\delta_{\omega,\mathbf{Z}}^k(\mathbf{X}, a, b)$ is calculated from the factors of the maximum variation of the observation function $\omega(\mathbf{Z}, \mathbf{X})$ and the length of the evaluation range $[a, b]$. The smaller the output value is, the better the resolution of a given function is achieved.

Although in real cases the optimal configuration cannot be achieved, it can be approximated by statistical reasoning of synthetic cases where the ground truth is known. Following this philosophy, an estimated $\hat{\delta}_\omega^k$

for each parameter k has been learned by averaging a small set of $\delta^k_{\omega,\mathbf{Z}}$. The learnt model is used for cases in which groundtruth is not available. By defining the ratio $\frac{\hat{\delta}^k_\omega}{\text{VAR}(S)^k}$ as the convergence criterion of the diffusion noise, a better estimation of the diffusion noise and, therefore, a more efficient sampling of the search space are achieved. In other words, the diffusion noise has to be reduced until the ratio converges to a value close to one, which indicates the resolution bound has been reached.

This reasoning has been incorporated into our new ArPF. This is done by changing Eq. (3) for the new Eq. (5), where the diffusion noise for the m-th layer is generated by $f^\sigma_m : \mathcal{R}^K \rightarrow \mathcal{R}^K$ proportionally to the initial noise $\mathbf{\Sigma} = \{\sigma_1, \cdots, \sigma_k\}$. The result is a semi-hard partitioning of the search space, where the set of functions F^σ promotes but not enforces the partitions for every level based on an annealed scheme of the noise. F^σ is expressed as M linear operators (being each function a $K \times K$ diagonal matrix), resulting into a annealed scheme defined by $M \times K$ different weights.

$$s^{(i)}_{k,m} = s^{(i)}_{k,m-1} + f^\sigma_m(\mathbf{\Sigma}), i = 0, \ldots, N-1 \qquad (5)$$

As mentioned before, these weights are learned from statistical reasoning from a set of synthetic experiments where only biomechanical constrains are used. The weights are constrained to fulfil two conditions. Firstly, the weights are selected following the order imposed by the hierarchy of the kinematic chain. Secondly, $\sigma_k = f^\sigma_m(k) > \cdots > f^\sigma_0(k) = \hat{\delta}^k_\omega$ is ensured the consistence of the simulated annealing, Eq. (1).

3. Experimental Results

In order to prove the advantages of our methodology, we apply ArPF in the context of hand tracking, as an example or hierarchical and complex articulated tracking. Our search space X is defined as a space of $D = 26$ dimensions (see Fig. 1) such as:

$$\mathbf{X} = \{x, y, z, \theta, \psi, \varphi, Y^1, Y^2, Y^3, Y^4, Y^5\} \qquad (6)$$

where x,y,z and θ,ψ,φ are the global translation and rotation parameters respectively, and $Y^i = \{\alpha, \beta_1, \beta_2, \beta_3\}$ are the angular vectors corresponding to each finger f_i.

The observation function is defined as a weighted combination of two terms, the hand visual appearance and its depth information (Eq. (7)), according to a given weight ρ, which in our case is 0.4.

$$h(\mathbf{X}, \mathbf{Z}) = \rho h_v + (1-\rho) h_d \qquad (7)$$

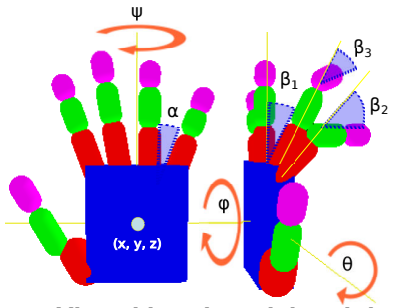

Figure 1. Virtual hand model and description of the state vector parameters.

The visual term is generated by the bi-directional silhouettes-based function $h_v(\mathbf{X}, \mathbf{Z})$ as given in [13]. The depth-based term is calculated as an extension of these silhouettes [13] for depth values:

$$h_d = \exp\left\{-\lambda \frac{\sum^{HxW} \textbf{thres}_{T1}(|D - \mathbf{Z}|)}{\sum^{HxW} \textbf{thres}_{T2}(D \times \mathbf{Z})}\right\} \qquad (8)$$

where λ is a constant to normalise the output values, and the operator "$\times$" represents the per-element multiplication of two matrices. The function **thres** is defined as $\textbf{thres}_{Ti}(\mathbf{X}) = 1$, if $X_j > Ti, \forall X_j \in X$; or X_j otherwise.

3.1. Quantitative evaluation

As additional contribution of this paper, we release a new dataset generator [12] of synthetic sequences and groundtruth for a detailed analysis of the problems inherent to articulated tracking. This is motivated for the lack of standard dataset and groundtruth in the particular field of hand tracking [10], as opposed to other application of such as HumanEva [13]. The tool provides to the community with a set of predefines 3D videos, the associated groundtruth and their correspondent 2D observations—silhouettes and depth maps—for comparison purposes. It also allows the researchers to configure and create new sequences and make them public by configuring individually the parameter values and variability of each of the segments in the output. Finally, the code of our tracking approach is also released and proposed as a baseline for comparison.

By using this dataset, the accuracy of ArPF is computed and compared with APF. Table 1 shows the results of ArPF and APF for five representative sequences that include translations (Seq.#1), rotations (Seq.#2) and complex fingers motion (Seq.#3-#5). The error metric applied here is the same described in [13] to calculate the average error of the articulations for full human body configurations. This metric allows for encoding all the angular parameters in a simple represen-

tation based on the 3D position of key points (the centroid of each finger articulation). Results consist on the average error of the key points according to the values provided by the groundtruth. All the methods where tested with 120 particles and 8 layers. APF was optimised for the parameter values, obtaining its optimal point for $k = 10$.

Table 1. Evaluation of the error [mm] for ArPF and different versions of APF.

Sequences	ArPF		APF [2]		APF [3]	
	mean	std.	mean	std.	mean	std.
#1	9	5	60	11	**7**	2.2
#2	7	3	65	19	**4**	1.8
#3	**7**	3.8	48	20	13	4.2
#4	**7**	2	60	14	11	3
#5	**8**	3	50	24	19	4.6

It can be observed how ArPF outperforms the best version of APF in 3 out of 5 cases. This is explained for the nature of the first 2 sequences, where no proper articulated motion is performed, but only global translation and rotation. In this context, APF is able to provide better results, but when complex hierarchical articulation movements appears (finger motion), ArPF improves APF clearly. This can be seen in Fig. 3, where both methods are tested for a periodic movement of opening and closing fingers (Seq.#3). For this case APF is not able of tracking such an action and just follows the global motion of the hand (its mass center and orientation). For this reason, APF generates an error that increases as the fingers move far away from their initial position. In contrast our approach is able to track the general motion and the complex movements successfully.

Our approach obtains an accuracy similar to the state of the art for hand tracking [10] by just applying our articulated tracking algorithm using simple observation functions and models.

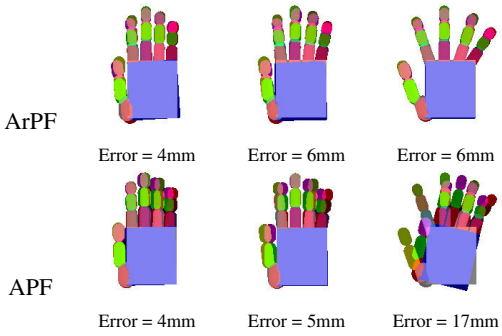

Figure 2. Pose estimation for ArPF and APF vs. the groundtruth (shadow) on sequence 3, frames: 1, 5, 20.

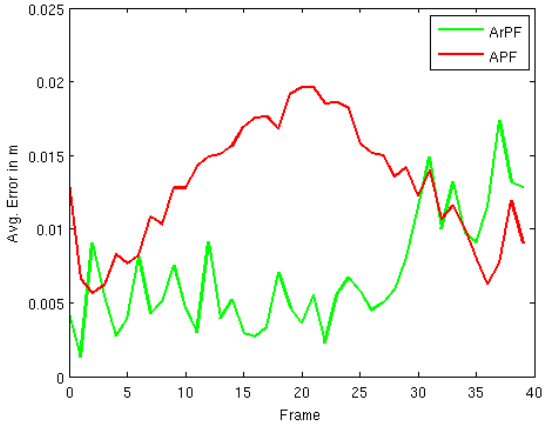

Figure 3. Evaluation of the error based on metric [13] for ArPF and APF [3] in seq. #3.

3.2. Qualitative evaluation

Finally, the performance of our algorithm is demonstrated on a real application, where Kinect is used as a sensor providing depth information and silhouettes are extracted by colour segmentation (a glove was used to simplified the segmentation since it is outside of the scope of this paper). Satisfactory results are shown in Fig. 4, where slight fitting errors are due to the simple modelling of the 3D hand.

Figure 4. Hand estimation for two real sequences capture by a Kinect.

4. Conclusions

In this paper a new version of particle filter for articulated tracking is presented. The ArPF makes use of a layered partitioning and a detailed noise modelling for sampling efficiently the search space. The potentiality of the method is demonstrated by using hand tracking as a case of use.

The main contributions of the paper are the formulation of the articulated particle filter, the extension of the concept of bidirectional silhouettes for depth maps and the introduction of a new dataset generator able to provide videos and groundtruth for testing quantitatively hand tracking algorithms [12]. Quantitative and qualitative results confirm the adequacy of our proposed methodology for articulated tracking. In spite of a rough 3D model of the hand and standard observation metrics, the system is able to successfully track the hand.

Acknowledgment

This work was supported by the Spanish MICINN, Plan E funds, under Grant TIN2009-14475-C04-02/01.

References

[1] F. Caillette, A. Galata, and T. Howard. Real-time 3-d human body tracking using learnt models of behavior. *CVIU*, 109(2):112–125, 2008.

[2] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, 2000.

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

[4] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *CVPR*, pages 73–80, 1996.

[5] Z. Husz, A. Wallace, and P. Green. Evaluation of a hierarchical partitioned particle filter with action primitives. In *EHuM*$_2$, 2007.

[6] S. Ivekovic, E. Trucco, and Y. Petillot. Human body pose estimation with particle swarm optimization. *Evolutionary Computation*, 16(4), 2008.

[7] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *PAMI*, 27:1805–1918, 2005.

[8] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV*, volume 2, pages 3–19, 2000.

[9] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *IJCV*, 53(3):199–223, 2003.

[10] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, pages 39–42, 2011.

[11] L. Raskin, M. Rudzsky, and E. Rivlin. Dimensionality reduction using a gaussian process annealed particle filter for tracking and classification of articulated body motions. *CVIU*, 2011.

[12] G. Ros and J. Martinez. Handy dataset, http://dipersec.king.ac.uk/handy/, (last accessed April 2012).

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

[14] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

[15] P. Wang and J. Rehg. A modular approach to the analysis and evaluation of article filters for figure tracking. In *CVPR*, volume 2, 2006.