LIKELIHOOD TUNING FOR PARTICLE FILTER IN VISUAL TRACKING

FONTMARTY Mathias†‡ , *LERASLE Frédéric*†‡, *DANÈS Patrick*†‡

† CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse, France
‡Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse, France

ABSTRACT

Particle filters (PF) are widely used in the Vision literature for visual object tracking. However, the selection and the tuning of the observation likelihood functions involved in the particle weighting stage are often eclipsed. These considerations have a strong influence on the tracking performance, especially for human motion capture (HMC) due to the high number of degrees of freedom and the presence of local minima in the state space. The proposed method is illustrated in the HMC context on a predefined set of likelihoods and assessed w.r.t. a ground truth provided by a commercial HMC system. This paper highlights the influence of their associated free parameters as well as their combination relevance in order to characterize the optimal unified likelihood function. These insights lead to some heuristics to tackle the difficult problem of the likelihood function tuning.

Index Terms— visual tracking, particle filtering, visual data fusion, tuning.

1. INTRODUCTION

The particle filtering framework [1], first introduced for visual tracking in the form of the CONDENSATION algorithm [2], has proved well suited for visual object tracking. The key idea is to represent the posterior distribution by a set of samples –or particles– with associated importance weights. This particle set is recursively updated over time taking into account the visual data by the means of the observation model. PFs make no restrictive assumption on the probability distributions entailed in the characterization of the problem, and permit a probabilistic principled fusion of diverse kinds of measurements. So, fusing multiple cues in an unified observation model is widely investigated in the Vision literature[3, 4]. This principle is somewhat questionable as the tuning of the free parameters involved in the likelihoods related to each visual cue is often omitted[1] and because the choice/relevance of the combined cues is seldom argued by quantitative evaluations. An empirical tuning method as well as derived heuristics are proposed here to optimize the unified likelihood with respect to visual tracking performances.

[1][5] is out of the paper scope as it discusses the tuning of the likelihood function when distinguishing tasks of single or multiple object view-based tracking, target loss recovery, etc.

Section 2 overviews our tracker implementation and the experimental setup for likelihood tuning. Among the amount of tracking applications, HMC is a good experimental field. Indeed, (i) appearance-based approaches are widely acknowledged as they offer a principled way to derive plethora of measurements [6, 7], (ii) the tracker fine tuning has a strong impact on the performance due to the presence of local minima which can trap the filter, (iii) a ground truth is available thanks to commercial HMC setup. Section 3 presents and comments the influence on the tracker performances when varying the free parameter values of the assumed observation model. These insights lead to the discussion and heuristics in section 4 to characterize and to fine tune the observation model given multiple cues with their associated likelihoods. Finally, section 5 summarizes our contributions and discusses future works.

2. FILTERING SCHEME AND EXPERIMENTAL SETUP

2.1. Framework

We formulate the tracking problem as the Bayesian state estimation of a Markovian stochastic process. In this context, we exploit the well-known CONDENSATION [2], which constitutes the basis algorithm on which more advanced ones are built. The estimated state is computed as the Minimum Mean Square Error Estimate (MMSEE). The human body is classically modelled by 9 truncated cones offering 22 degrees of freedom (DOF). Its configuration parameters to be estimated at time k constitute the hidden state vector $\mathbf{x}_k$ of a system which delivers the measurements –images– $\mathbf{z}_k$. This system is characterized by its state dynamics $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ –here a random walk– and the conditional density $p(\mathbf{z}_k|\mathbf{x}_k)$ of the output given the state vector, which also defines the likelihood of the state w.r.t. the measurements. These measurements are inferred from the human body model projection in trinocular image sequences.

The experimental setup involves three IEEE 1394 "progressive scan" Flea 2 color cameras providing 640×480 images. Ground truth positions of the template joints are given by a commercial HMC system from Motion Analysis. Both systems are software calibrated and synchronized. In order to analyze the average filter performance, 30 runs are performed on each set of data. We assess the performance on various se-

quences of $\sim 20\ s$ including walking, arm waving, pointing, and fitness.

2.2. Metrics design for filter evaluation

Several criteria are defined, entailing the MMSEE delivered by each run of the filter and the ground truth from the commercial system. As comparing joint angles may be tricky, and as significatively different values of the state vector can lead to similar aspects of the 3D template, the four following metrics rely on the true positions of the joint centers of our human body model, and their estimates provided by the filter:

RMS error w.r.t. true joint center positions - The first criterion is the average on all joints of the RMS errors between the joint center estimates and ground truth computed over all frames and filter runs. This criterion, henceforth termed "RMSE" is fairly standard in the Vision literature [8, 7, 9].

Bias - This criterion checks if multiple runs of the filters provide an estimate centered on the ground truth.

Estimator dispersion - The variance of the HMC system is analyzed over the filter runs, to ensure that given an input video stream, the computed estimates are stable enough despite their stochastic nature.

Tracking failure rate - To complete the evaluation, a last criterion focuses on the failure rate. The tracker is considered to fail each time the distance between any true and estimated joint center position exceeds a certain threshold function of the bias.

2.3. Observation likelihoods

Our tracker implementation considers as examples four visual cues whose associated likelihoods have the following form *i.e.*

$$p(\mathbf{z}_k^{c,i}|\mathbf{x}_k) \propto \exp\left(-\frac{D_{c,i}^2}{2\sigma_i^2}\right), \qquad (1)$$

where $D_{c,i}$ terms the similarity distance for the c-th camera and the i-th cue, and σ_i is its standard deviation beforehand determined.

Silhouette-based likelihood (sil) - In the vein of [3], we first perform a foreground-background silhouette segmentation in order to obtain the silhouette mask. We sample points inside the limbs of each projected particle to check whether or not the limbs are consistent with the segmented silhouette.

Dual silhouette-based likelihood (sil2) - In order to complete this first cue, we counterbalance it by the one proposed in [6]. The principle is symmetric, consisting in sampling the segmented silhouette and check its consistency w.r.t. the current projected particle.

Skin color-based likelihood (*skin*) - To improve localization accuracy—especially for thin limbs such as arms—, we set up an additional likelihood function involving skin color computation. A skin color probability image is computed thanks to the backprojection of an off-line learnt skin color histogram. For each hypothesis $\mathbf{x}_k$, we compute the average skin color probability on 3 virtual points situated on the projected head and hands of the current particle. A high average skin color probability on these 3 points results in a low similarity distance for the current hypothesis.

Skin blob distance-based likelihood (*skin_dist*) - We also propose a variant of the preceding cue. Skin color blobs are extracted from the skin color probability image. For each projected particle, we compute the distance between head and hands and the nearest detected skin blob. The best configurations are the ones showing the lowest distances.

Values of the similarity distances may be different from one to the others. Thus, in our implementation, $D_{sil} \in [0, 255], D_{sil2} \in [0, 1], D_{skin} \in [0, 255], D_{skin\_dist} \in [0, 50]$. The four likelihoods have a form similar to (1) provided that σ_{sil}, σ_{sil2}, σ_{skin}, σ_{skin_dist} term the corresponding standard deviations. Assuming they are mutually independent conditioned on the state $\mathbf{x}_k$, the unified likelihood factorizes as :

$$p(\mathbf{z}_k|\mathbf{x}_k) = \prod_{c=1}^{C} \prod_{i} p(\mathbf{z}_k^{c,i}|\mathbf{x}_k).$$

with $i \in \{sil, sil2, skin, skin\_dist\}$. Figure 1 illustrates the influence of such free-parameters in the unified likelihood on the tracking process. Depending on the σ_i parameter values, we observe a tracking failure of the left arm in this sequence.

Fig. 1. Tracking on a simple sequence with $(\sigma_{sil}, \sigma_{sil2}, \sigma_{skin}, \sigma_{skin\_dist}) = (130, 0.5, 20, 30)$ –top– and $(\sigma_{sil}, \sigma_{sil2}, \sigma_{skin}, \sigma_{skin\_dist}) = (20, 0.1, 60, 5)$ –bottom–. The superimposed avatar represents the estimated configuration.

3. EXPERIMENTS AND RESULTS

Measurement design and combination can greatly affect the filter behavior. This is why we assess the effect of the σ_i parameter in equation (1).

3.1. Measurement design

Figure 2 presents the RMSE of the estimated template for our 2 measurement cues $skin$ and $skin\_dist$, quite close in

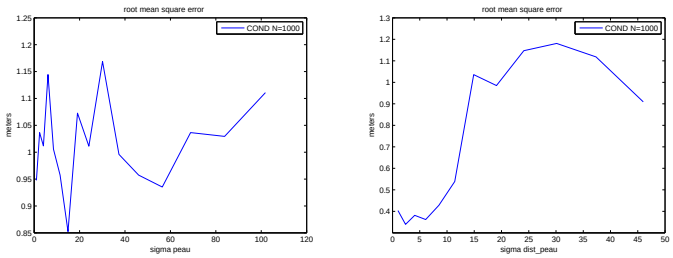


Fig. 2. Influence of σ_{sil} on RMSE for 30 runs of the CONDENSATION filter when $skin$ cue is used alone (left) and when $skin\_dist$ is used alone (right).

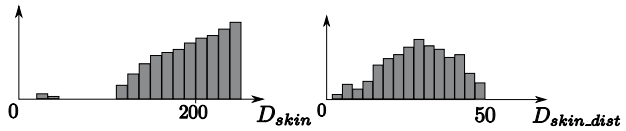


Fig. 3. Histograms of the similarity distances D_{skin} (left) and D_{skin_dist} (right) for a given particle cloud.

their definition. Others metrics are not presented due to lack of space but they present the same characteristics: $skin$ cue present catastrophic results (RMSE $\sim$ $1m$) in comparison with $skin\_dist$ cue, whatever the chosen σ_{skin}. This is due to its very peaky nature. Indeed, as high probability skin regions may be very small, head and hand of the projected particles are rarely exactly situated in a high probability region (low D value) as can be seen on figure 3 - left. Thus, their weight is systematically flattened. On the other side, the $skin\_dist$ cue presents a smooth profile due to the distance computation, what enables a more regular distribution of similarity distances (cf. figure 3 - right). This shows how important the design of the likelihood functions is.

3.2. Measurement tuning

As the sil cue is the most used in the literature, we first focus on the tracker behavior when used alone in the composite likelihood. Figure 4 (blue curve) presents the influence of the σ_{sil} parameter on a simple tracking with $N = 1000$. First of all, we can notice that the optimal value of σ_{sil} depends on the criterion to be minimized. RMSE leads to $\sigma_{sil} = 22$, dispersion tends to be minimum for $\sigma_{sil} = 200$, and optimal bias is reached for $\sigma_{sil} = 1$. Below this value, the tracking diverges due to computational underflows (near 0 calculations). Intuitively, a high value of σ_{sil} results in a less spread estimate as the considered cue is given less importance with respect to the assumed prior dynamics. Visually, this corresponds to a smoother evolution from one frame to the other. RMSE and bias are lower for a lower σ_{sil}, however, the minimum bias is $0.1\ m$ per joint and the optimal RMSE is $0.17\ m$, what reveals that the single sil cue is not sufficient to perform a satisfying tracking with 1000 particles. Obviously, a trade-off has to be made between accuracy and dispersion.

In order to obtain a good compromise, we chose $\sigma_{sil} = 30$. We proceed similarly by including in the composite likelihood the $sil2$ and $skin\_dist$ cues, what leads to $\sigma_{sil2} = 0.07$ and then $\sigma_{skin\_dist} = 5$. This last cue finally enables a satisfying tracking and lowers the RMSE to $0.06\ m$. However, one can wonder if sil and $sil2$ cues are not redundant. We present on figure 4 (red curve) the σ_{sil} variation effect while using sil, $sil2$ and $skin\_dist$ cues together. The filter behavior for high values of σ_{sil} informs us about the irrelevance of the sil cue, as the filter tends to perform as if it was not exploited. Thus, we notice that the sil cues slightly improves the RMSE, but none of the other criteria. We can draw two conclusions here: first, some cues may be superfluous to use in combination with others and just only add an unnecessary computational load without improving global results. Second, this behavior witnesses a correlation between measurements. This suggests that the property of independence between measurements conditionally to the state may be partially inappropriate.

4. DISCUSSION AND GENERAL HEURISTICS

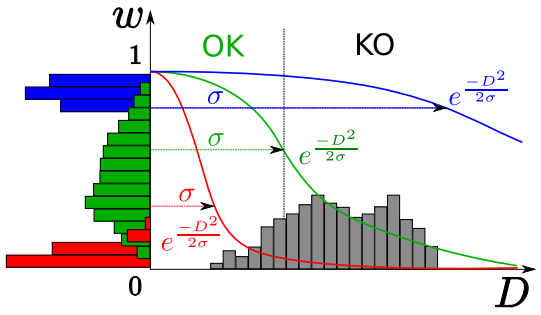
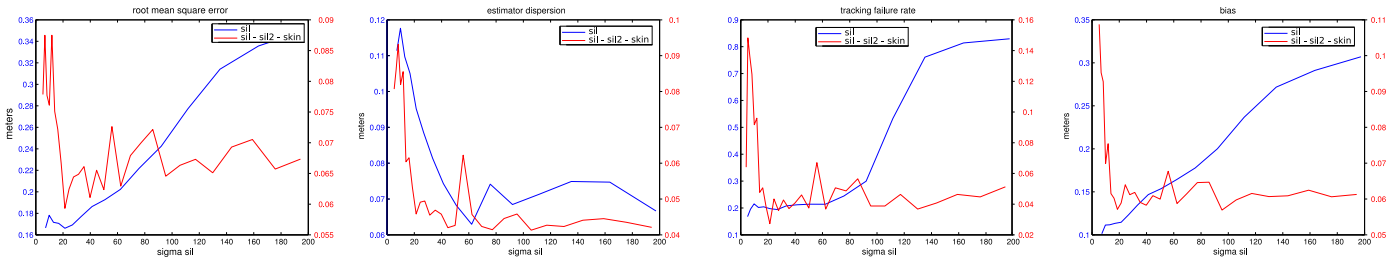


Fig. 5. Histogram of similarity distances for a particle cloud *vs.* histogram of the associated weigths for different σ choices.

From our evaluations, we can derive a few observations concerning the tuning of the likelihoods: (1) there is a trade-off between accuracy and dispersion, (2) a low $\sigma_{(.)}$ value enables a lower bias and a greater mean accuracy, (3) too many measurements/cameras involved or too low $\sigma_{(.)}$ values lead to a greater dispersion and in the worst case to filter divergence due to numerical problems (underflows). Thus, we propose the following heuristics to guide us in the tuning of the likelihoods:

Smooth similarity distance - First of all, one must choose a similarity distance function with a smooth profile in order to enable a well-balance distribution of their values for a given particle cloud (grey histogram in figure 5). One can note that a null distance similariy (*i.e.* a perfect match between particle and visual data) never occurs due to our imperfect models (human, environments, state-measurement link).

Early σ value tuning - The evolution of this distance w.r.t. the state vector entries can constitute a good guideline to the tuning. Figure 5 presents the similarity distance histogram (grey) for a given particle cloud. The early tuning of σ must be situated within the range of the histogram. Basically, particles with a shorter similarity distance than the chosen σ value



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

Fig. 4. Influence of σ_{sil} for 30 runs of the CONDENSATION filter depending on whether sil cue is used alone (blue) or together with $sil2$ and $skin$ cues (red) : (a) RMSE, (b) estimator dispersion, (c) failure rate, (d) bias.

(green) are considered as "good" ones w.r.t. the considered cue. The others are considered too "bad", and they will be affected a low weight. This can be supported by the fact that the highest slope of a gaussian $\mathcal{N}(x;\mu,\sigma^2)$ is reached for $x=\sigma$.

Refining the values - We can sharpen the initial choice according to our observations: if we augment σ (blue curve in figure 5), all weights will be near 1 and the considered cue brings less information. Thus it will be less taken into account in the filtered density, what will lower the dispersion of the estimate. If we decrease σ, the likelihood function shows a higher peak, which results in a more drastic selection of the fittest particles (red curve and histogram on figure 5). Meanwhile, the estimator will show a high dispersion on various runs. Moreover, values of the computed weights globally decrease too, what can lead to the following problem.

Computational limits - The value of σ cannot be decreased at will : indeed, a too low value results in a null likelihood function due to computer encoding limits (*e.g.* $\exp(-\frac{1}{2}\frac{0.5^2}{0.01^2})=0$, so that selecting $\sigma=0.01$ results in null weights for a similarity distance of 0.5 and leads to the filter failure). This phenomenon is amplified when we use multiple cues and/or multiple cameras as likelihood functions are multiplied. Consequently, one has to be very cautious at decreasing σ values and mixing various heterogeneous cues.

5. CONCLUSION AND FUTURE WORKS

We proposed a study of likelihood tuning in a visual human motion capture context with respect to four criteria. It appears that mixing measures is not trivial and the likelihood tuning has a strong influence on these criteria. Compromises have to be made between the tracking accuracy and dispersion, which has been seldom mentioned in the literature. Adding more measurements can also significantly impact on the filter dispersion and may not improve global behavior while needlessly increasing computational load. From these insights, we propose some simple heuristics to estimate the range of the involved free parameters, namely the standard deviations.

Some interesting investigation lines from this work could involve more advanced measurements. In addition, to complete this study, one should also add visual detectors and importance sampling methods to enable automatic initialization.

6. REFERENCES

[1] A. Doucet, N. De Freitas, and N. J. Gordon, *Sequential Monte Carlo Methods in Practice*, Series Statistics For Engineering and Information Science. Springer-Verlag, New York, 2001. 1

[2] M. Isard and A. Blake, "CONDENSATION – Conditional density propagation for visual tracking," *International Journal on Computer Vision (IJCV'98)*, vol. 29, no. 1, pp. 5–28, 1998. 1

[3] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision (IJCV'05)*, vol. 21, no. 3, pp. 185–205, 2005. 1, 2

[4] H. Moon and R. Chellappa, "3D shape-encoded particle filter for object tracking and its application to human body tracking," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2008. 1

[5] J. Lichtenauer, Reinders M. J. T., and Hendriks E. A., "Influence of the observation likelihood function on particle filtering performance in tracking applications," in *Automatic Face and Gesture Recognition (FGR'04)*, Seoul, KOREA, May 2004, pp. 767–772. 1

[6] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *International Journal on Robotic Research (IJRR'03)*, vol. 6, no. 22, pp. 371–393, May 2003. 1, 2

[7] P. Wang and J. Rehg, "A modular approach to the analysis and evaluation of particle filters for figure tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, USA, 2006, pp. 790–797. 1, 2

[8] A. Balan, L. Sigal, and M. Black, "A quantitative evaluation of video-based 3D person tracking," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05)*, Washington, USA, October 2005, pp. 349–356. 2

[9] "http://vision.cs.brown.edu/humaneva/ — humaneva," . 2