Bootstrapping Sequential Monte Carlo Tracking

Thomas B. Moeslund and Erik Granum

Laboratory of Computer Vision and Media Technology
Aalborg University, Denmark
`{tbm,cg}@cvmt.dk`

Abstract. Sequential Monte Carlo (SMC) methods have in recent years been applied to handle some of the problems inherent to model-based tracking. In this paper we suggest to apply bootstrapping to reduce the required number of particles in SMC tracking. By bootstrapping is meant to track reliable low-level image features and use them to bootstrap the high-level model-based tracking. The concept of bootstrapped SMC tracking is exemplified by monocular tracking of the 3D pose of a human arm with the position of the hand in the image as the bootstrapping information. Tests suggest that both bootstrapping is a sound strategies and an improvement over standard SMC-methods.

1 Introduction

In model-based tracking a high dimensional solution space often appears and an exhaustive search is therefore seldom practical. The standard solution to this problem is to use a prediction followed by either an iterative search, a Kalman Filter, or an exhaustive search in the proximity of the prediction. The drawback of these approaches is the risk of ending up in a local extremum, i.e. estimating the wrong state. In recent years, statistical methods such the Condensation algorithm [6][14], the particle filter [11][13], and Multi-Hypothesis tracking [2][3] have therefore been applied to approximate an exhaustive search, or in statistical terms - approximate the posterior probability density function (PDF). These methods all belong to the class of Sequential Monte Carlo (SMC) methods [5].

A Monte Carlo method represents the posterior PDF by a finite number of weighted state samples (known as particles) each selected from an Importance Function and weighted by the measurements. This sampling principle is known as Importance Sampling. An SMC method is a Monte Carlo method operating on a temporal sequence of measurements. Here the Importance Function can be defined by predicting the posterior PDF from the previous time instant. In other words, each of the most likely states in the posterior PDF in the previous time instant is sampled, predicted into the current time instant, and compared to the current measurements in order to obtain a weight. The weight reflects the similarity between the predicted state and the measurements, i.e. the likelihood. The predicted states and their associated weights approximates the posterior PDF in the current time instant. The current state of the system is defined as the maximum a posterior PDF (MAP).

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 1030–1037, 2003.

A common problem for SMC methods is, the dependency on a good prediction. If the prediction is accurate, fewer particles are required to estimate the posterior PDF. If, on the other hand, the prediction is less precise, an artificially high process noise is required to allow the particles to diffuse in the state-space and estimate the posterior PDF [7]. The reason for not having an accurate prediction is that it is difficult to model the dynamics in a particular tracking scenario, due to complex motions and lack of ground truth information.

1.1 Outline of the Paper

In this paper, we will try to improve the prediction and thereby reduce N. This will be based on a bootstrapping approach.

In section 2 we define bootstrapping in general and show how it can be applied to improve the quality of the prediction. The following sub-sections are devoted to exemplifying this idea in the context of monocular tracking of the 3D pose of a human arm utilising an SMC-approach. In section 3 we present our results and discuss our findings, and in section 4 a conclusion is given.

2 Prediction and Bootstrapping

Predicting a state, $\overrightarrow{U}(t)$, from time $t-1$ to time t is usually done by adding a deterministic part, $\overrightarrow{D}(T)$, and a stochastic part, $\overrightarrow{S}(T)$, hence $\overrightarrow{U}(t) = \overrightarrow{D}(T) + \overrightarrow{S}(T)$, where T indicates dependencies on the entire past, i.e. $T = \{0, 1, ..., t-1\}$. $\overrightarrow{D}(T)$ consists of a motion model, $\overrightarrow{M}(T)$, which describes how the state evolves over time. $\overrightarrow{M}(T)$ contains a number of parameters whose current values are kept in $\overrightarrow{\omega}(T)$. $\overrightarrow{M}(T)$ is usually independent on time. $\overrightarrow{\omega}(T)$ is typically estimated in a recursive framework where it is assumed to be a first order Markov process, hence $\overrightarrow{\omega}(T) = \overrightarrow{\omega}(t-1)$. In practise the deterministic part is normally defined as $\overrightarrow{D}(T) = \overrightarrow{D}(\overrightarrow{M}, \overrightarrow{\omega}(t-1))$ [1].

The stochastic part is added to model the errors in the motion model and is referred to as the process noise. $\overrightarrow{S}(T) = \overrightarrow{S}(\overrightarrow{N}(T)\ ,\ \overrightarrow{\phi}(T))$ where $\overrightarrow{N}(T)$ is the model of the process noise and $\overrightarrow{\phi}(T)$ is the current values of the parameters in this model. The process noise is often assumed to be independent on time and modelled as a Gaussian distribution. And as above a first order Markov process is assumed, hence $\overrightarrow{\phi}(T) = \overrightarrow{\phi}(t-1)$. In practise the stochastic part is therefore normally defined as $\overrightarrow{S}(T) = \overrightarrow{S}(\overrightarrow{N}, \overrightarrow{\phi}(t-1))$, where $\overrightarrow{N}$ is a multivariate Gaussian distribution.

2.1 Bootstrapping

When tracking an object it is sometimes possible to recognise parts of the object prior to tracking. For example, in the context of tracking the 3D human figure in a monocular image sequence it is in general difficult to find robust features to track. However, some features can actually be tracked independent on others. These are: the face/head, the hands, the feet, and in some cases also other distinct points, e.g. arm pits, shoulders, and crotch.

Say we are able to find one of these features, denoted by $\vec{\beta}(t)$. This would allow a comparison between $\vec{\beta}(t)$ and $\vec{U}(t)$ resulting in an estimate of (parts of) the prediction error, i.e. $\vec{\phi}(t)$. Applying $\vec{\phi}(t)$ as opposed to $\vec{\phi}(t-1)$ obviously gives a far better estimate of the stochastic part. We denote the new estimate with a plus, $\vec{S}(T)^+ = \vec{S}(\vec{N}, \vec{\phi}(t))$.

Furthermore, $\vec{\beta}(t)$ also contains information that can be used to bias the deterministic prediction, or more precisely $\vec{\beta}(t)$ can correct (parts of) the predicted deterministic part. That is, given $\vec{\beta}(t)$ we can estimate $\vec{\omega}(t)$ and apply this instead of $\vec{\omega}(t-1)$. We denote the new estimate of the deterministic part as $\vec{D}(T)^+ = \vec{D}(\vec{M}, \vec{\omega}(t))$. So, instead of predictions based on estimates at time $t-1$ we now use our estimates from time t, $\vec{\beta}(t)$, to correct our predictions, altogether providing a better result.

We denote this approach *bootstrapped tracking*. The success of this approach depends on how much information is carried in $\vec{\beta}(t)$, i.e. how many of the state's parameters can be corrected, and how much this information can prune the state-space representation.

2.2 Bootstrapping the State-Space Representation

A concrete tracking problem is required in order to evaluate the idea of bootstrapped tracking. We use the problem of monocular tracking of the 3D pose of a human arm as a case study.

In this work we find the position of the hand in the images, $[h_x, h_y]^T$, using colour segmentation [9] and let this be our bootstrapping information, hence $\vec{\beta}(t) = [h_x, h_y]^T$. In the context of $\vec{\beta}(t)$ a geometric model of the arm needs to be defined. Before doing so we introduce the following assumptions: the hand is a part of the lower arm, the 3D position of the shoulder is known, and the lengths of the upper arm (A_u) and lower arm (A_l) are known. We derive a compact representation of the arm through our bootstrapping information $\vec{\beta}(t)$ together with the screw axis representation [12].

In the context of modelling the human arm the screw axis is defined as the vector spanned by the shoulder and the hand, and the position of the elbow is defined by rotating an initial elbow position, α, degrees around the screw axis [9]. Combining $\vec{\beta}(t)$ with the camera parameters obtained during calibration, the position of the hand in the image can be mapped to a line, l, in space passing through the hand. That is, one parameter is sufficient to represent the 3D position of the hand. We denote this parameter H_z. Combining H_z with α we have a two-parameter state-space representation as oppose to the standard four parameters applied to represent the arm (Euler's angles) [9]. So we have bootstrapped the state-space representation resulting in a reduced size of the state-space, and hence, a reduced N.

2.3 Bootstrapping the SMC-Algorithm

SMC is defined in terms of Bayes' rule and using the first order Markov assumption. That is, the posterior PDF is equal to the observation PDF multiplied by the prior PDF, where the prior PDF is the predicted posterior PDF from time $t-1$:

$$p(\overrightarrow{X_t}|\overrightarrow{\theta_t}) = p(\overrightarrow{\theta_t}|\overrightarrow{X_t})p(\overrightarrow{X_t}|\overrightarrow{\theta_{t-1}}) \tag{1}$$

where $\overrightarrow{X}$ is the state, i.e. $\overrightarrow{X} = [\alpha, H_z]^T$ and $\overrightarrow{\theta}$ is the image measurements. The predicted posterior PDF is defined as

$$p(\overrightarrow{X_t}|\overrightarrow{\theta_{t-1}}) = \int p(\overrightarrow{X_t}|\overrightarrow{X_{t-1}})p(\overrightarrow{X_{t-1}}|\overrightarrow{\theta_{t-1}})\ \mathrm{d}\overrightarrow{X_{t-1}} \tag{2}$$

where $p(\overrightarrow{X_t}|\overrightarrow{X_{t-1}})$ is the motion model governing the dynamics of the tracking process, i.e. the prediction, and $p(\overrightarrow{X_{t-1}}|\overrightarrow{\theta_{t-1}})$ is the posterior PDF from the previous frame. As described in section 1 SMC estimates $p(\overrightarrow{X_t}|\overrightarrow{\theta_t})$ by selecting a number, N, of (hopefully) representative states (particles) from $p(\overrightarrow{X_{t-1}}|\overrightarrow{\theta_{t-1}})$, predicting these using $p(\overrightarrow{X_t}|\overrightarrow{X_{t-1}})$, and finally given each particle a weight, π, in accordance with the measurements $p(\overrightarrow{\theta_t}|\overrightarrow{X_t})$. So, as explained earlier, a key issue is to have an accurate prediction.

To apply bootstrapping to the SMC-algorithm we need to apply $\beta(t)$ in order to define $\overrightarrow{D}(T)^+$ and $\overrightarrow{S}(T)^+$. At this point it might be in order to emphasise that our state-space representation is in the two parameters α and H_z, but all calculations are done in the anatomic representation, i.e. the 3D position of the elbow, $\overrightarrow{E}$, and the 3D position of the hand, $\overrightarrow{H}$. These two representations co-exist and their relationship is given via Rodriques' formula [4]. We apply bootstrapping first for the position of the hand, $\overrightarrow{H}$, and then for the position of the elbow, $\overrightarrow{E}$.

The correction of the prediction of $\overrightarrow{H}$ is based on the idea of combining the prediction and the image measurements, $\beta(t)$. In figure 1 the predictions are illustrated using subscript 'p' while the corrected predictions are illustrated using subscript 'c'.

Since we know the camera ray through the hand in the current image, l, we can correct the prediction by projecting the predicted position of the hand, $\overrightarrow{H_p}$, to the line, l. The corrected prediction is denoted $\overrightarrow{H_1}$ and calculated as $\overrightarrow{H_1} = \overrightarrow{P} + ((\overrightarrow{H_p} - \overrightarrow{P}) \cdot \overrightarrow{F})\overrightarrow{F}$ where $\overrightarrow{P}$ is the focal point and $\overrightarrow{F}$ is the unit direction vector for the line l. The stochastic prediction models the process noise by diffusing the deterministic prediction. As we know the hand is on the line, l, we diffuse it by randomly sampling from a Gaussian distribution located along the line, l, see figure 1. The mean of the Gaussian is defined by $\overrightarrow{H_1}$ and the standard deviation controlled by the error vector, i.e. standard deviation $= c_1 \cdot ||\overrightarrow{H_p} - \overrightarrow{H_1}||$ where c_1 is a predefined constant. After this operation we have the final corrected position of the hand, $\overrightarrow{H_c}$. The difference between the predicted and corrected

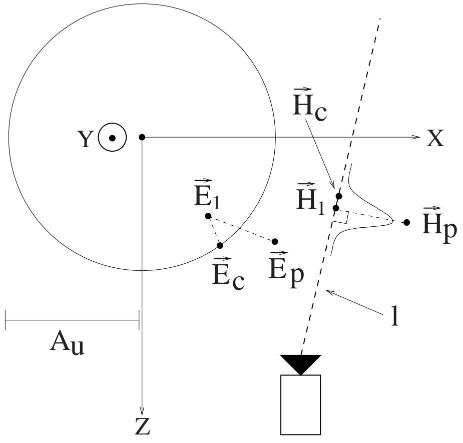


Fig. 1. The shoulder coordinate system seen from above. The circle illustrates the sphere limiting the position of the elbow. The dotted line indicates a camera ray through the hand. See text for a definition of the parameters.

vectors yields a measure of the prediction error (innovation), denoted $\overrightarrow{H_e}$ and calculated as $\overrightarrow{H_e} = \overrightarrow{H_c} - \overrightarrow{H_p}$.

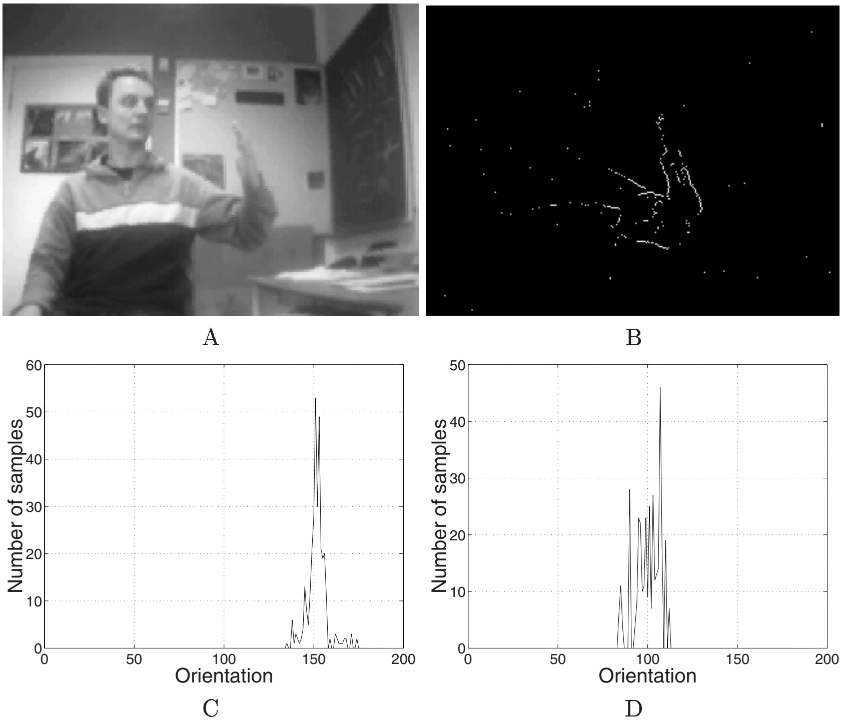
The predicted position of the elbow can not directly be bootstrapped by $\overrightarrow{\beta}(t)$. However, we know it is likely to have a predicted error closely related to that of the hand as the hand and elbow are parts of the same kinematic chain. We therefore calculate the corrected position, $\overrightarrow{E_c}$, by first adding the predicted error of the hand to the predicted value of the elbow, yielding $\overrightarrow{E_1} = \overrightarrow{E_p} + \overrightarrow{H_e}$, and then finding the point closest to $\overrightarrow{E_1}$ that results in a legal configuration of the arm. In mathematical terms $\overrightarrow{E_c} = arg \min_{\overrightarrow{E}} \left|\left|\overrightarrow{E} - \overrightarrow{E_1}\right|\right|$ subjected to the constraints $\left|\left|\overrightarrow{E}\right|\right| = A_u$ and $\left|\left|\overrightarrow{EH_c}\right|\right| = A_l$. The solution to this problem can be found in [10].

Evidently, the prior in the SMC-approach will be much more accurate when applying the bootstrapping, and this is true even with a very simple motion model.

2.4 Defining the Observation PDF

The observation PDF needs to be defined in order to apply all of the above to actually estimate the 3D pose of the human arm. That is, we need to define the representation of the image and how to compare that to the state-space representation, i.e. how to calculate the weight, π.

Our image representation is in the form of the orientations of the upper arm and lower arm, respectively. In figure 2.A a typical input image is shown. We first find all temporal edges, shown in figure 2.B, by ANDing an edge image



A B

C D

Fig. 2. A: A typical input image (shown in B/W). B: The temporal edge pixels. C: The distribution of the orientation of the upper arm. D: The distribution of the orientation of the lower arm.

and a difference image. We have the positions of the shoulder, hand, and can estimate the position of the elbow utilising prediction. This allows us to divide all temporal edge pixels into two groups, one for the upper arm and one for the lower [8]. In each group a variant of the Hough Transform is applied to estimate the orientations of the two arm parts [8]. In the figures 2.C and 2.D the estimated orientation of figure 2.A are shown. We view the two figures as PDFs, denoted $p(\theta_u)$ and $p(\theta_l)$, and therefore normalise them so they each sum to 1. This complex image representation clearly allows for a simple comparison. That is, we project a state-space configuration (α, H_z) into the image, calculating the two orientations, θ_u and θ_l, and finally calculating the weight as $\pi = p(\theta_u) + p(\theta_l)$.

3 Results

In our implementation of the bootstrapped SMC-approach we tried different values of N. In some cases N can be chosen as low as 10 and still producing a good estimate of the MAP. In general $N = 50$ produces good results, but sometimes up to 100 samples are required to insure a good approximation of the posterior PDF and the MAP. We chose $N = 100$.

A test is conducted to illustrate the effect of utilising bootstrapping compared to traditional SMC. In the case of standard SMC the tracker was manually initialised to the correct pose 100 frames earlier than the image shown in figure 2. After 100 frames the five best MAPs are illustrated in figure 3 with- and without bootstrapping. The five best MAPs are illustrated in both a 3D plot and projected into the image. Figure 3 clearly shows the superior performance of our approach compared to standard SMC tracking.

In images such as the one in figure 2.A the posterior PDF is in general ambiguous. In this particular case, a number of correct poses can be found by increasing α as the distance between the hand and camera increases. This tendency can be seen in figure 3 and it means that the estimated MAP might be incorrect in this particular image. However, due to the ill-posed nature of the problem this will always be the case independent on the chosen tracking framework. The good thing is that in this tracking framework *the* correct state is virtually always among the largest peaks and is therefore very likely to evolve into the next image. That is, the bootstrapped tracking approach handles multiple hypotheses.

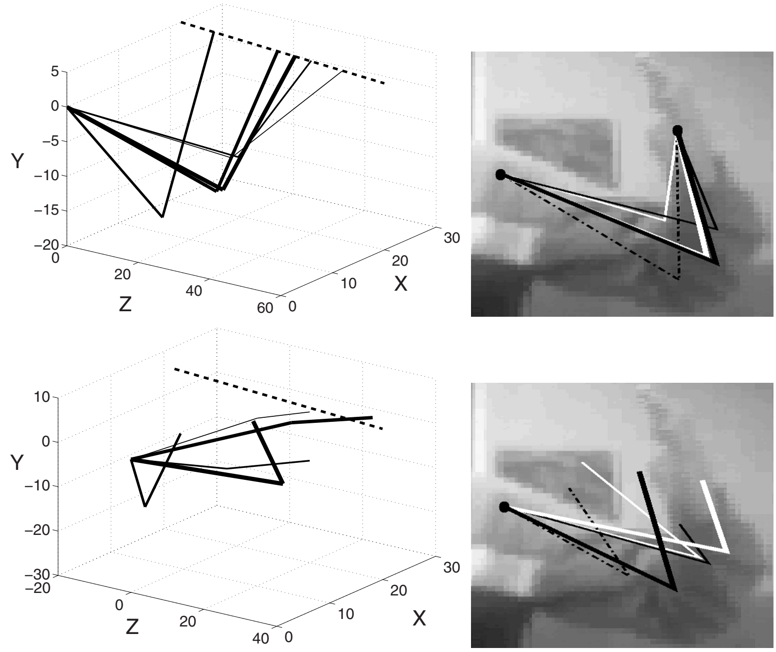


Fig. 3. The five most likely configurations of the arm in the image in figure 2.A with bootstrapping (first two) and without (last two) bootstrapping. For the 3D plots: the thicker the line the higher the likelihood. The dotted line illustrates the camera ray passing through the hand. For the 3D configurations projected into the image: the probability of the lines are in the following order (smallest probability first): thin white, thin black, dash-dotted, thick white, thick black).

4 Conclusion

In this paper we have suggested to apply bootstrapping to increase performance in model-based tracking. Besides the tests presented above the improvement achieved by this approach can also be understood intuitively. Just imagine the complex nature of the posterior PDF having utilised four Euler angles instead of our two, α and H_z. Concretely our primary contribution is the idea of applying image measurements from the current frame to improve the state-space representation, the predictions, and the process noise. We have illustrated the effects of this contribution in the context of an SMC method, but the idea of bootstrapped tracking is also valid in other tracking frameworks.

References

1. A. Blake and M. Isard. *Active Contours.* Springer, 1998.
2. T.J. Cham and J.M. Rehg. A Multiple Hypothesis Approach to Figure Tracking. In *CVPR*, Fort Collins, Colorado, USA, June 23-25 1999.
3. Y. Chen, Y. Rui, and T. Huang. Mode-based Multi-Hypothesis Head Tracking Using Parametric Contours. In *International Conference on Automatic Face and Gesture Recognition*, Washington D.C., USA, May 20-21 2002.
4. J.J. Craig. *Introduction to Robotics. Mechanics and Control.* Addison Wesley, second edition, 1989.
5. A. Doucet, N. Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.
6. M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal on Computer Vision*, pages 5–28, 1998.
7. M. Isard and A. Blake. ICONDENSATION: Unifying Low-level and High-level Tracking in a Stochastic Framework. In *ECCV*, Freiburg, Germany, June 2-6 1998.
8. T.B. Moeslund. Improving Sequential Monte Carlo Tracking by Bootstrapping. Technical Report CVMT 02-02, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2002.
9. T.B. Moeslund and E. Granum. Modelling and estimating the pose of a human arm. *Machine Vision and Applications.* To appear.
10. T.B. Moeslund and E. Granum. Modelling the 3D Pose of a Human Arm and the Shoulder Complex utilising only Two Parameters. Submitted to *Conference on Model-based Imaging, Rendering, image Analysis and Graphical special Effects, March 2003, France.*
11. H. Sidenbladh, M.J. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *ECCV*, Copenhagen, Denmark, 2002.
12. V.M. Zatsiorsky. *Kinematics of Human Motion.* Champaign, IL: Human Kinetics, 1998.
13. Z. Zeng and S. Ma. Head Tracking by Active Particle Filtering. In *International Conference on Automatic Face and Gesture Recognition*, Washington D.C., USA, May 20-21 2002.
14. S. Zhou, V. Krueger, and R. Chellappa. Face Recognition from Video: A CONDENSATION Approach. In *International Conference on Automatic Face and Gesture Recognition*, Washington D.C., USA, May 20-21 2002.