RUN-TIME DYNAMIC MODEL ESTIMATION FOR A MULTI-HYPOTHESIS MEAN SHIFT TRACKER

Tom Caljon and Peter Schelkens

Vrije Universiteit Brussel - Institute for Broadband Technology
Department of Electronics and Information Processing
Brussels, Belgium.

ABSTRACT

We resort to a mean-shifting color-based particle filter for the purpose of tracking objects in a video sequence. Primary use case is the creation of interactive versions of video sequences i.e. video sequences containing clickable objects. As the tool needs to be used by non-technical users specification of parameters should be kept to a minimum. For objects types that were not anticipated and for which a trained dynamical model is not available, we then propose run-time estimation of the dynamical model instead of simpler models such as constant velocity with fixed noise variance. We present results on the performance of such tracker and compare with non-meanshifting particle filters and a single-hypothesis mean shift tracker.

1. INTRODUCTION

Reliable tracking of an object in a video sequence has been a research topic for several decennia. As generic vision systems with capabilities comparable to those of a human are not to be expected soon, current trackers are application-specific. Their design choices depend on a number of constraints : complexity of the proposed algorithm (realtime vs non-realtime), availability of measurements (online vs offline processing), representation of an object's state (e.g. rectangle, parameters of a deformable template, contour, bitmask, ...), inclusion of uncertainty in the results (single-hypothesis, single hypothesis with uncertainty, multi-hypothesis), robustness (e.g. to lighting changes, occlusion, rotation, ...), discriminating features of the target (color, texture, sound, ...), world constraints (fixed vs moving camera, known vs unknown scene, ...) etc.

This work should be seen in the context of a tracker that follows objects in video sequences available offline. Intended use cases are the addition of interactivity to video sequences and region of interest coding. Objects are represented by their bounding box in the frame. The input material is not subjected to specific requirements, so the camera may be fixed or moving and the scene may be unknown. The main idea is that the user can track any type of object and has to set only a limited number of parameters. Examples of existing algorithms suited for these conditions are color based mean shift trackers [4], color based particle filters [8], and the online appearance modelling technique by Jepson *et al* [5].

We experiment with a multi-hypothesis mean shift tracker, hoping that much less particles are needed than a regular particle filter to obtain good results. As mean shift drives each initial hypothesis (rectangle) to the mode of the likelihood and not the posterior distribution, this operation should be applied with care. The initial hypotheses have to be well chosen and the number of mean shift iterations constrained. To accomodate for objects with unknown dynamics, we perform online dynamical model estimation. Section 2 and 3 describe the multi-hypothesis mean shift tracker, section 4 the run-time dynamical model estimation. Results are presented in section 5.

2. STATE SPACE MODEL

For robustness reasons, in tracking the estimate for the unknown state $\hat{x}_t$ is often balanced between the prediction by a dynamical model and the measurement z_t given by the observation model at timestep t, using their respective uncertainties (noise). Existing frameworks for such estimation are Kalman filters [6] and particle filters [1]. Kalman filters require the observation and prediction noise to be Gaussian, a constraint dropped by particle filters at the cost of added complexity: a plethora of samples in object state space with associated weights is used to approximate probability densities. The weighted sample set approximation of the posterior distribution $p(x_t|z_{1:t})$ can be filtered to the next timestep given the state transition prior $p(x_t|x_{t-1})$ and a likelihood distribution $p(z_t|x_t)$. In this work, the states $x_t = (x, y, w, h)^T$ are rectangles with components x, y (center of the rectangle), width w and height h. $p(x_k|x_{k-1})$ can be a distribution corresponding to a first-order auto-regressive process

$$x_k = A_1 x_{k-1} + d + Bw \tag{1}$$

or a second-order auto-regressive process

$$x_k = A_2 x_{k-2} + A_1 x_{k-1} + d + Bw \tag{2}$$

This requires specification of A_1 and A_2 (4-by-4), d (4-by-1, constant drift) and B (4-by-4, uncertainty, responsible for spreading the standard normal noise in the components of w). For the likelihood $p(z_k|x_k)$ we use the function

$$exp(-20D[q^*, q_k(x_k)]^2) \tag{3}$$

from [8], where $D[q^*, q_k(x_k)]$ is the Bhattacharrya distance [4] between the model HSV histogram q^* and $q_k(x_k)$, the HSV histogram of frame k's pixels inside x_k.

3. MULTI-HYPOTHESIS MEAN SHIFT

The proposed tracker combines particle filters[1] and the mean shift tracking algorithm[4] as previously done in [9]. Each posterior density $p(x_t|z_{1:t})$ is approximated by N weighted hypotheses

$$\{((x_{t|t}^{(n)}, \ldots, x_{1|t}^{(n)}), \pi_t^{(n)})|n \in \{1, \ldots, N\}\}.$$

Each hypothesis is a sequence of states (path), where $x_{m|t}^{(n)}$ is the state at timestep m that propagated out of the state path $(x_{m-1|t}^{(n)}, \ldots, x_{1|t}^{(n)})$. The weight $\pi_t^{(n)}$ indicates the belief at timestep t that $x_{t|t}^{(n)}$ is the real state x_t of the object, and $\sum_{n=1}^{N} \pi_t^{(n)} = 1$.

As is usually done for particle filters, the importance density from which samples are taken is the state transition prior $p(x_k|x_{k-1})$.

However, prior to weighting the samples using the likelihood $p(z_k|x_k)$, each sample is allowed to evolve towards the local modes of the likelihood distribution using mean shift. Only a couple of mean shift iterations are performed, in order to prevent overly ignoring of the transition prior and collapsing of all samples onto one point (or just a couple of points) in object state space.

The algorithm can be summarized as

At timestep $t+1$ do:

1. Resample systematically N samples from $p(x_t|z_{1:t})$ i.e. select N hypotheses

$$\{(x_{t|t}^{(n)'}, \ldots, x_{1|t}^{(n)'})|n \in \{1, \ldots, N\}\}$$

with

$$p((x_{t|t}^{(n)'}, \ldots, x_{1|t}^{(n)'}) = (x_{t|t}^{(j)}, \ldots, x_{1|t}^{(j)})) = \pi_t^{(j)}$$

2. $\forall i \in \{1, \ldots, N\}$
   (a) Predict: sample w from a multi-variate standard normal distribution and set

$$x_{t+1}^{(i)} = A_2 x_{t-1|t}^{(i)'} + A_1 x_{t|t}^{(i)'} + d + Bw$$

   or likewise for a first-order model using eq 1.
   (b) Meanshift $x_{t+1}^{(i)}$ for m iterations
   (c) Calculate weight: $\pi_{t+1}^{(i)} = p(z_{t+1}|x_{t+1}^{(i)})$
   (d) $h_{t+1}^{(i)} = (x_{t+1}^{(i)}, x_{t|t}^{(i)'}, \ldots, x_{1|t}^{(i)'})$
3. Normalize weights: $\pi_{t+1}^{(i)} = \frac{\pi_{t+1}^{(i)}}{\sum_{j=1}^{N} \pi_{t+1}^{(j)}}$
4. The weighted samples approximation of $p(x_{t+1}|z_{1:t+1})$ is then $\{(h_{t+1}^{(i)}, \pi_{t+1}^{(i)})|i \in \{1, \ldots, N\}\}$

A handy user interface allows for initialization and correction of tracked objects, specification of the dynamical models and mean shift parameters. The required model histograms q^* can be loaded or calculated from pixels inside the initial object rectangles.

4. RUN-TIME DYNAMICAL MODEL ESTIMATION

We want to be capable of tracking an object whose dynamics are unknown. Rather than using a default, fixed auto-regressive dynamical model with a large B that highly spreads the predictions, we try to learn the dynamics as tracking goes on. At the start of the tracking process, $p(x_k|x_{k-1})$ is initialised for a broad distribution from which ample initial hypotheses are drawn e.g. a distribution corresponding to a first order auto-regressive model $x_k = Ax_{k-1} + Bw$, with $A = I$ and B a diagonal matrix

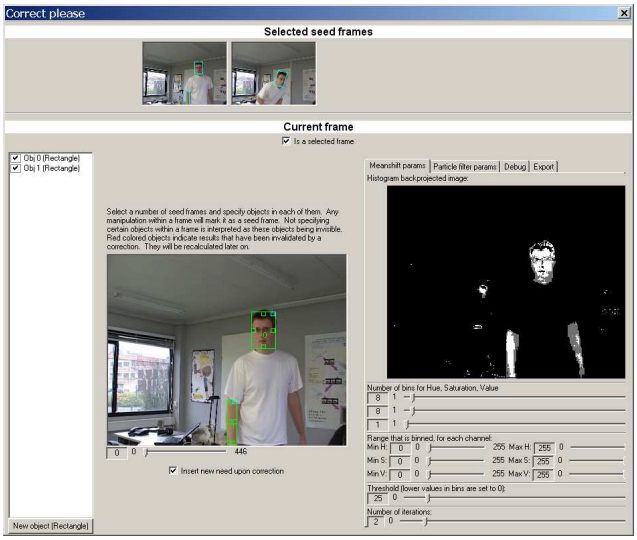

Fig. 1. User interface for initialization and correction.

that sufficiently spreads the independent standard normal noise in each component of w, or a constant velocity model with fixed noise variance. However, once enough state posterior distributions $\{p(x_k|z_{1:k})|k = 1 \ldots m\}$ are available, a more accurate model may be trained. We propose to estimate A_1, A_2, $\bar{x}$ and B of a second order auto-regressive model

$$x_k - \bar{x} = A_2(x_{k-2} - \bar{x}) + A_1(x_{k-1} - \bar{x}) + Bw$$

(where $\bar{x}$ represents the mean target state) at each timestep $t > m$ and to use this model for the prediction of $x_t^{(1)}, \ldots, x_t^{(N)}$. The estimation at timestep t makes use of the tracker's posterior distributions $\{p(x_j|z_{1:j})|j = t - W \ldots t - 1\}$. When W is a constant, training is performed using a sliding window over the posteriors, when $W = t - 1$ all available posteriors are used.

We first performed some offline tests on ground truth data $x_1^*, \ldots, x_M^*$ for a frantically moving red stick using maximum likelihood estimation (MLE). The maximum likelihood estimates for the parameters of a second-order auto-regressive model parameters can be calculated using moments R_i and autocorrelations R_{ij} of ground truth using these equations [2]:

$$R_i = \sum_{k=3}^{M} x_{k-i}^*, R_{ij} = \sum_{k=3}^{M} x_{k-i}^* (x_{k-j}^*)^t$$

$$R'_{ij} = R_{ij} - \frac{1}{M-2} R_i R_j^t$$

$$A_2 = (R'_{02} - R'_{01}R'^{-1}_{11}R'_{12})(R'_{22} - R'_{21}R'^{-1}_{11}R'_{12})^{-1}$$

$$A_1 = (R'_{01} - A_2 R'_{21})R'^{-1}_{11}$$

$$D = \frac{1}{M-2}(R_0 - A_2 R_2 - A_1 R_1)$$

$$\bar{x} = (I - A_2 - A_1)^{-1} D$$

$$B = \sqrt{\frac{1}{M-2}(R_{00} - A_2 R_{20} - A_1 R_{10} - D R_0^t)}$$

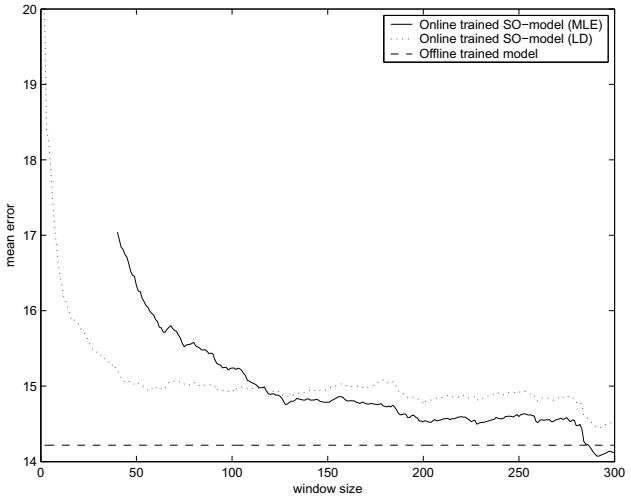

Fig. 2. Mean of prediction errors at each timestep, in function of number of past states used to perform MLE of A_1, A_2, B, $\bar{x}$ at that timestep

For small window sizes invertibility problems arose, so the Levinson-Durbin (LD) algorithm [3] that does not involve matrix inversions was also used. Figure 2 shows the mean prediction error

$$e = \frac{1}{M}\sum_{i=1}^{M} \|x_i^* - (A_2(x_{i-2}^* - \bar{x}) + A_1(x_{i-1}^* - \bar{x}) + \bar{x})\| \quad (4)$$

in function of W, as well as the mean prediction error when using a model trained offline from $x_1^*, \ldots, x_M^*$. As expected, when W is too small, estimation uses uncharacteristic (biased) data, causing a larger error. However, some models obtained from large windows perform better than the globally trained model, which is possible when the object behaves differently in parts of the sequence. For this sequence, models trained using $W > 100$ already perform well.

Switching from offline training using ground truth to online training from filtered estimates of the real states introduces the possibility that the model is trained using incorrect posteriors, which will cause a bad dynamical model. In spite of this vulnerability we proceed with the estimation of $\varphi_i = (A_{1,i}, A_{2,i}, \bar{x}_i, B_i)$ from $\{p(x_j|z_{1:j})|j = i-W, \ldots, i-1\}$ at timestep i. We assume a sliding window of size W and allow $\varphi_i \neq \varphi_j$. North [7] shows how to derive plausible n-th order auto-regressive models from weigthed samples posteriors offline. It is an expectation maximization (EM) technique that in the E-step of the $(l+1)$-th iteration estimates (using $A_{1,i}^{(l)}$, $A_{2,i}^{(l)}$, $\bar{x}_i^{(l)}$ and $B_i^{(l)}$) the sufficient statistics R_{ij} and R_i $(i, j \in \{1, 2\})$ for the maximum likelihood estimation of $A_{1,i}^{(l+1)}$, $A_{2,i}^{(l+1)}$, $\bar{x}_i^{(l+1)}$ and $B_i^{(l+1)}$ in the M-step:

1. Choose suitable values for the starting dynamics
2. Run the Condensation tracker on the training image sequence using the current dynamics, producing a sample-set representation of the distributions $p(x_t|z_{1:t})$
3. (Optionally but advised) run the Condensation smoothing algorithm, to obtain smoothed distributions $p(x_t|z_{1:T})$ where T is the last timestep
4. E-step: find expected values for the moments R_i and autocorrelations R_{ij} :

$$E[R_{ij}] \approx \sum_{n=1}^{N} \pi_T^{(n)} \sum_{t=3}^{T} x_{t-i|T}^{(n)} (x_{t-j|T}^{n})^t$$

$$E[R_i] \approx \sum_{n=1}^{N} \pi_T^{(n)} \sum_{t=3}^{T} x_{t-i|T}^{(n)}$$

5. M-step: estimate (MLE) the dynamical model from these expected values
6. Goto step (2)

For online use however, retracking the target within the considered window several times per timestep would be intractable. So we simply stick to estimating the moments and autocorrelations just once given the newly obtained posterior $p(x_{i-1}|z_{i-1})$ and discarding $p(x_{i-W-1}|z_{1:i-W-1})$, followed by maximum likelihood estimation of φ_i:

At each timestep i, estimate the second order model $\varphi_i = (A_{1,i}, A_{2,i}, \bar{x}_i, B_i)$ used for sampling $x_i^{(1)}, \ldots, x_i^{(N)}$ as follows:

1. E-step: find expected values for the moments R_i and autocorrelations R_{ij} from $p(x_{i-1}|z_{1:i-1}) = \{((x_{i-1|i-1}^{(n)}, \ldots, x_{1|i-1}^{(n)}), \pi_{i-1}^{n}), n = 1, \ldots, N\}$:

$$E[R_{lk}] \approx \sum_{n=1}^{N} \pi_{i-1}^{(n)} \sum_{t=i-W+2}^{i-1} x_{t-l|i-1}^{(n)} (x_{t-k|i-1}^{n})^t$$

$$E[R_l] \approx \sum_{n=1}^{N} \pi_{i-1}^{(n)} \sum_{t=i-W+2}^{i-1} x_{t-l|i-1}^{(n)}$$

   This can be done fast by reusing calculations of the previous timestep.
2. M-step: estimate (MLE) the dynamical model from these expected values

5. RESULTS

We tested the tracker on two sequences: the frantically moving red stick sequence and the wandering man sequence. The first sequence contains 766 frames of a fast moving stick with many direction changes, the second (447 frames) is more relaxed. Distractors were not included, as they would require a smarter like-

lihood function that deals better with occlusions, clutter, lighting changes, etc. Ground truth was obtained by running the proposed tracker and correcting by hand. Following trackers were tested on each sequence:

- static second-order model, trained offline from ground truth, 3 mean shift iterations, 20 hypotheses (PF+MS).
- tracking using a second-order model, estimated at run-time, with $W = 30$, $W = 60$, $W = 120$, 3 mean shift iterations , 20 hypotheses (PF+MS+O).
- no mean shift, offline trained second-order model, 500 particles (PF).
- single-hypothesis mean shift using a constant velocity model with fixed variance ($B_{ii} = 3$, $B_{ij} = 0 (i \neq j)$) (MS).
- same as above, but 20 hypotheses (PF+MS+F).

Tables 1 and 2 present the following results: the mean (1) and standard deviation (2) over time t of $\|\hat{x}_t - x_t^*\|$, the mean (3) and standard deviation (4) over time t of the mean distance between non-shifted hypotheses and ground truth $\frac{1}{N}\sum_{i=1}^{N}\left\|x_t^{(i)} - x_t^*\right\|$ and the mean (5) and standard deviation (6) over time t of the smallest distance between a non-shifted hypothesis and ground truth $min(\{\left\|x_t^{(i)} - x_t^*\right\| | i \in \{1, \ldots, N\}\})$. Results were collected by running each tracker 4 times over each sequence, averaging the obtained figures and rounding them towards the nearest integer.

Tracker	(1)	(2)	(3)	(4)	(5)	(6)
PF+MS	4	9	22	9	9	7
PF+MS+O, W=30	12	46	33	48	18	40
PF+MS+O, W=60	5	9	22	10	10	9
PF+MS+O, W=120	4	10	23	11	11	10
PF	73	37	97	33	17	14
MS	10	9	24	12	24	12
PF+MS+F	5	10	21	14	15	14

Table 1. Results for the red stick sequence (see text).

Tracker	(1)	(2)	(3)	(4)	(5)	(6)
PF+MS	6	3	7	3	5	3
PF+MS+O, W=30	6	3	7	3	5	3
PF+MS+O, W=60	6	3	7	3	5	3
PF+MS+O, W=120	6	3	8	3	5	3
PF	29	15	32	15	13	10
MS	6	3	9	3	9	3
PF+MS+F	7	5	9	4	5	4

Table 2. Results for the wandering man sequence (see text)

From (5) and (6) it seems the model trained offline from ground truth is most succesful at predicting the location of the target in the next frame. However, the difference in prediction and overall tracking performance is small and does not even exist for the easier second sequence. Too small window sizes lead to the expected increase in error (W=30 in table 1). Even with 500 particles, PF performs significantly less than PF+MS(+O). Although the best predictions lie relatively close to the true state, the histogram-based likelihood does not prefer these good predictions to smaller rectangles lying within the object. Mean shift does not suffer from this problem due to an expanding window, which however can cause problems when similar-colored clutter is present in the background. In such situations, additional features may be used in the likelihood. MS scores less well in prediction than PF+MS(+O). The same goes for PF+MS+F, but both recover for these sequences.

The tested particle filters operate at about 1 frame per second (on a 1 Ghz processor, CIF resolution), on average 10 times slower than the multi-hypothesis mean-shift trackers.

6. CONCLUSION

We presented our multi-hypothesis mean shift tracker and confirmed the good performance claimed in [9]. Additionally, we proposed run-time estimation of a second order dynamical model for objects whose dynamics are unknown. The technique was tested and performed well for the considered sequences given a reasonable amount of initial tracking results gathered using an initial constant velocity model.

7. ACKNOWLEDGMENTS

This work is a result of the Advanced Media project, a joint collaboration between the Vrije Universiteit Brussel, VRT, IMEC and Universiteit Gent. Peter Schelkens holds a post-doctoral fund with the Fund for Scientific Research Flanders (FWO).

8. REFERENCES

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):173–188, 2002.

[2] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, 1998.

[3] P. Brockwell and R. Dahlhaus. Generalized levinson-durbin and burg algorithms. *Journal Of Econometrics*, 118(1-2):129–149, 2004.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 25(5):564–577, May 2003.

[5] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 25(10):1296–1311, 2000.

[6] R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(D):35–45, 1960.

[7] B. North. *Learning Dynamical Models for Visual Tracking*. PhD thesis, University of Oxford, 1998.

[8] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, pages 661–675, 2002.

[9] C. Shan, Y. Wei, T. Tan, and F. Ojardias. Real time hand tracking by combining particle filtering and mean shift. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 669–674, 2004.