Robust Multiple-People Tracking Using Colour-Based Particle Filters

Daniel Rowe [1], Ivan Huerta [1], Jordi Gonzàlez [2], and Juan J. Villanueva [1]

[1] Computer Vision Centre, Universitat Autònoma de Barcelona, Spain
[2] Institut de Robòtica i Informàtica Industrial, UPC, Barcelona, Spain

Abstract. Robust and accurate people tracking is a key task in many promising computer-vision applications. One must deal with non-rigid targets in open-world scenarios, whose shape and appearance evolve over time. Targets may interact, causing partial or complete occlusions. This paper improves tracking by means of particle filtering, where occlusions are handled considering the target's predicted trajectories. Model drift is tackled by careful updating, based on the history ofl ikelihood measures. A colour-based likelihood, computed from histogram similarity, is used. Experiments are carried out using se quences from the CAVIAR database.

1 Introduction

Robust and accurate people tracking is a key task in many promising computer-vision applications, such as smart video surveillance or human-computer interfaces [6,4,1]. The interest in multiple-people tracking is also prompted by the challenge of emulating the amazing capabilities of natural systems to detect motion and keep lock on several moving objects simultaneously.

However, serious difficulties should b e expected. The system must deal with non-rigid targets, often highly articulated and elastic, who may wear loose-fitting clothes. In open-world applications, neither the number of targets, nor their appearance or shape can be specified in advance. Considerable foreground diversity should be taken into account. Further, both observed shape and appearance evolve over time depending on the point of view, or on the local illumination and background, specially if these are uncontrolled. Finally, as the targets interact, they may group and split, causing occlusions, and changing the observed appearance and shape. This paper enhances tracking by means of particle filtering (PF). A preliminary work was published in [11]. The main contributions of the presented approach are the following:

- it copes with clutter distracters by adopting a colour-based likelihood computed from histogram similarity. Colour information relative to the target surroundings is used to tune the colour histograms.
- It deals with multiple targets simultaneously, paying special attention to the sampling impoverishment phenomenon. The systems scales well with the number of targets, avoiding the curse of dimensionality common to PF.

- Model drift is precluded by careful updating, based on likelihood measures, thereby ensuring proper tracking despite noisy measures, estimate errors, occlusions, and changes in illumination conditions and camera viewpoint.
- Occlusions are handled considering the predicted trajectories of all targets within the scene and the history of likelihood measurements.

The remainder of this paper is organised as follows. Section 2 covers the probabilistic framework and related approaches. A colour-based particle filter for multiple-target tracking is proposed in Section 3. Section 4 shows some experimental results, and section 5 summarises the conclusions.

2 Probabilistic Tracking Framework

The computation of the state $\mathbf{s}_t$ given all evidence to date $\mathbf{e}_{1:t}$ is called *filtering*. The posterior pdf $p(\mathbf{s}_t \mid \mathbf{e}_{1:t})$ can be calculated through *recursive estimation*:

$$p(\mathbf{s}_t \mid \mathbf{e}_{1:t}) \propto \underbrace{\underbrace{p(\mathbf{e}_t \mid \mathbf{s}_t)}_{\text{likelihood}}}_{\text{updating}} \underbrace{\int \underbrace{p(\mathbf{s}_t \mid \mathbf{s}_{t-1})}_{\text{trans. model}} \underbrace{p(\mathbf{s}_{t-1} \mid \mathbf{e}_{1:t-1})}_{\text{previous post.}} d\mathbf{s}_{t-1}}_{\text{prediction}}. \tag{1}$$

The pdf is projected forward according to the transition model, making a prediction. It is then updated in agreement with the new evidence, $\mathbf{e}_t$. When non-Gaussian, non-linear distributions are involved, this problem is overcome by simulating N i.i.d. random samples from the posterior pdf, $\left\{\mathbf{s}_t^i; i = 1:N\right\}$. This leads to the *particle filter approach*. This works as follows: the posterior pdf at time $t-1$, $p(\mathbf{s}_{t-1} \mid \mathbf{e}_{1:t-1})$, is represented by a weighted set of samples, $\left\{\hat{\mathbf{s}}_{t-1}^i, \overline{\pi}_{t-1}^i; i = 1:N\right\}$. The set is re-sampled using normalised weights $\overline{\pi}_{t-1}^i$ as probabilities. The temporal prior$\left\{\hat{\mathbf{s}}_t^i\right\}$ is obtained by applying the transition model $p(\mathbf{s}_t \mid \mathbf{s}_{t-1})$ to each sample. The likelihood $p(\mathbf{e}_t \mid \mathbf{s}_t)$ is represented by weights π_t^i, which are then normalised. Expectations are approximated as:

$$\mathbb{E}_{p(\mathbf{s}_t \mid \mathbf{e}_{1:t})}(\mathbf{s}_t) \simeq \sum_{i=1}^{N} \overline{\pi}_t^i \hat{\mathbf{s}}_t^i. \tag{2}$$

Although SIR methods have been widely used in recent years, they have important drawbacks [7]. *Sampling impoverishment* is one of the main ones: samples are spread around several *modes* pointing out hypotheses in the state space, but most of them may be spurious. Unfortunately, there is a non-negligible probability of losing modes, a low probability of recovering them and the remaining modes could be all spurious. Different approaches have been taken in order to overcome these and other issues. Nummiaro et al. [9] use a PF based on colour-histogram cues. However, no multiple-target tracking is considered, which implies that no scene event such as target grouping or occlusion can be analysed. Perez et al. [10] propose also a PF based on a colour-histogram likelihood. They introduce interesting extensions in multiple-part modelling, incorporation of background information, and multiple-target tracking. Nevertheless, it requires an extremely

large number of samples, since one sample contains information about the state of all targets, dramatically increasing the state dimensionality. Further, no appearance model updating is performed, what usually leads to target loss in dynamic scenes. Deutscher et al. [3] present an interesting approach called *annealing particle filter* which aims to reduce the required number of samples. However, pruning hypotheses with lower likelihood could be undesirable in a cluttered environment. Contour tracking have also been explored [8], although this may be inappropriate if used as the only cue in crowded scenarios because of multiple occlusions. BraMBLe [5] is an appealing approach to multiple-blob tracking which models both background and foreground using Mixtures of Gaussians (MoG). However, no model update is performed, there is a common foreground model for all targets, and suffers for the curse of dimensionality, since it tackles multiple-target tracking combining information about all targets in every sample. Therefore, even though a great number of improvements have been introduced in recent years, there is still much ground to cover.

3 A Multi-target Colour-Based PF

The motion of the central point of an elliptical region is modelled using first-order dynamics in image coordinates. The l-labelled target's state is defined as $\mathbf{s}_t^l = \left(\mathbf{x}_t^l, \mathbf{u}_t^l, \mathbf{w}_t^l, \mathbf{q}_t^l, \rho_t^l, \lambda_t^l\right)^T$, where components are the ellipse position, velocity, both axes, the appearance model, the occlusion status, and the expected target likelihood. A label l associates a specific appearance model to the corresponding samples, allowing multiple-target tracking. Given the high dimensionality of images, a feature extraction process is mandatory. In this approach, evidences $\mathbf{e}_t$ are given by colour histograms computed at each predicted location and size.

After the initialisation, no sample is generated using detection, since it would mask tracking misbehaviours. Thus, just tracking performances are tested by means of propagating hypotheses and weighting them according to evidence. Clearly, by incorporing detection, the general performance will be enhanced, providing the system with error-recovery capabilities.

3.1 Transition Model

The position, speed, and size of each sample are predicted according to:

$$\begin{aligned} \hat{\mathbf{x}}_t^{i,l} &= \mathbf{x}_{t-1}^{i,l} + \mathbf{u}_{t-1}^{i,l}\Delta_t + \xi_{\mathbf{x}}^i, \\ \hat{\mathbf{u}}_t^{i,l} &= \mathbf{u}_{t-1}^{l} + \xi_{\mathbf{u}}^i, \\ \hat{\mathbf{w}}_t^{i,l} &= \mathbf{w}_{t-1}^{i,l} + \xi_{\mathbf{w}}^i. \end{aligned} \tag{3}$$

The random vectors $\xi_{\mathbf{x}}^i, \xi_{\mathbf{u}}^i, \xi_{\mathbf{w}}^i$, sampled from WAGN processes, provide the system with a diversity of hypotheses. Sample likelihoods depend on sample position and size, but not on their speeds. Thus, if speeds were propagated considering the previous speed, they would be in quasi open loop[1]. Thus, their values

[1] There would still be a weak relation, since speeds are used to predict positions, and position errors can be measured, but a considerable delay would be introduced.

could become completely different from the true values within a few frames, and an important proportion of samples would be wasted. In order to avoid this phenomenon, the estimated target speed $\mathbf{u}_{t-1}^{l}$ at time $t-1$ is fed back into the prediction of $\hat{\mathbf{x}}_t^{i,l}$.

3.2 Likelihood Function

The likelihood function computes the pdf of image features given the state. The target appearance can be represented by means of colour histograms. Histograms are broadly used to represented human appearance, since they are claimed to be less sensitive than colour templates to rotations in depth, the camera point of view, non-rigid targets, and partial occlusions. Thus, the $l-$model is given by:

$$\mathbf{q}^{l} = \left\{ q_k^{l}; k = 1 : K \right\}, \tag{4}$$

where K is the number of bins, and the probability of each feature is:

$$q_k^{l} = C^{l} \sum_{a=1}^{M} \delta \left(b \left(\mathbf{x}_a \right) - k \right), \tag{5}$$

where C^{l} is a normalisation constant required to ensure that $\sum_{k=1}^{K} q_k^{l} = 1$, δ the Kronecker delta, $\{\mathbf{x}_a; a = 1 : M\}$ the pixel locations, and $b\left(\mathbf{x}_a\right)$ a function that associates the given pixel to its corresponding histogram bin. The target distribution at the predicted position $\hat{\mathbf{x}}_t^{i,l}$ and ellipse size $\hat{\mathbf{w}}_t^{i,l}$, is given by $\mathbf{p}_i^{l}$, which is calculated in the same way as the model. The similarity between two histograms can be computed using the following metric [2,9]:

$$d_B = \sqrt{1 - \rho \left(\mathbf{p}, \mathbf{q} \right)}, \tag{6}$$

where $\rho \left(\mathbf{p}, \mathbf{q} \right) = \sum_{k=1}^{K} \sqrt{p_k q_k}$. is the *Bhattacharyya coefficient.* Therefore, similar histograms have a high Bhattacharyya coefficient, which should correspond to high sample weights. The computed metric can be mapped using a Gaussian distribution [9], and samples are thus weighted according to:

$$\pi_t^{i,l} = p\left(\mathbf{e}_t \mid \hat{\mathbf{s}}_t^{i,l} \right) = \mathcal{N}\left(d_B; \mu, \sigma^2 \right). \tag{7}$$

So far no background information has been used. However, tracking success depends on how distinguishable the target is from a local environment. Thus, foreground features present also in its surroundings should be less important for target localisation. Here, an approach similar to [2] is adopted by using a *centre-surround* model to compute the background histogram $\mathbf{r}^{l}$ according to the outer region which encloses the target. Hence, the background histogram is used to compute a weight for each bin:

$$\omega_k = \left\{ \min \left(\frac{r_k^{*}}{r_k} \right); k = 1 : K \right\}, \tag{8}$$

where r_k^{*} is the minimum non-zero value. Thus, these weights are then applied to both model and target histograms to diminish the importance of those bins which represent the local background.

3.3 Weight Normalisation

In a multiple-target tracking scenario, those targets whose samples exhibit lower likelihood are more likely to be lost, since the probability of propagating one mode is proportional to the cumulative weights of its samples. In order to avoid one target absorbing other target samples, genetic drift must be prevented. Thus, a memory term, which takes into account the number of targets being tracked, is included. Weights are normalised according to:

$$\overline{\pi}_t^{i,l} = \frac{\pi_t^{i,l}}{\sum_{i=1,j=l}^{N} \pi_t^{i,j}} \frac{1}{L}, \tag{9}$$

where L is the number of tracked targets. Each weight is normalised according to the total weight of the target's samples. Thus, all targets have the same probability of being propagated, since the addition of the weights of each target samples sums $\frac{1}{L}$. This allows multiple-target tracking using a single PF, despite the differences between their likelihoods and the genetic drift phenomenon.

3.4 State Estimation

The l-target estimates are computed according to:

$$\begin{aligned}
\mathbf{x}_t^l &= (1-\alpha_{\mathbf{x}}) \left(\mathbf{x}_{t-1}^l + \mathbf{u}_{t-1}^l \Delta_t \right) + \alpha_{\mathbf{x}} \left(L \textstyle\sum_{i=1}^{N} \overline{\pi}_t^{i,l} \widehat{\mathbf{x}}_t^{i,l} \right), \\
\mathbf{u}_t^l &= (1-\alpha_{\mathbf{u}}) \, \mathbf{u}_{t-1}^l + \alpha_{\mathbf{u}} \left(\frac{\mathbf{x}_t^l - \mathbf{x}_{t-1}^l}{\Delta_t} \right), \\
\mathbf{w}_t^l &= (1-\alpha_{\mathbf{w}}) \, \mathbf{w}_{t-1}^l + \alpha_{\mathbf{w}} \left(L \textstyle\sum_{i=1}^{N} \overline{\pi}_t^{i,l} \widehat{\mathbf{w}}_t^{i,l} \right),
\end{aligned} \tag{10}$$

where $\alpha_{\mathbf{x}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{w}} \in [0,1]$ denote the adaptation rates. Target speeds are not estimated according to sample speeds and their weights, since significant errors would be introduced: samples are chosen only because of sample weights, which do not directly depend on the current speed. This fact could imply a significant amount of jitter and many samples would be wasted. Therefore, target speeds are computed from successive position estimates. Further, both position and speed estimates are enhanced by regularising them according to their histories.

The target appearance must also be updated. However, this is a sensitive task which may lead to the well-known *model drift* phenomenon. Thus, models are then only updated when two conditions hold: (i) the target is not occluded and (ii) the likelihood of the estimated target's state suggests that the estimate is sufficiently reliable. In this case, they are updated using an adaptive filter:

$$\mathbf{q}_t^l = (1-\alpha_{\mathbf{q}}) \, \mathbf{q}_{t-1}^l + \alpha_{\mathbf{q}} \mathbf{p}_t^l, \tag{11}$$

where $\alpha_{\mathbf{q}} \in [0,1]$ is the learning rate. In order to determine when the estimate is reliable, the likelihood of the current estimate is computed, $p\left(\mathbf{e}_t \mid \mathbf{s}_t^l\right)$.

The appearance is then updated when this value is higher than an indicator of the expected likelihood value, calculated following an adaptive rule:

$$\lambda_t^l = (1 - \alpha_l)\, \lambda_{t-1}^l + \alpha_l p\left(\mathbf{e}_t \mid \mathbf{s}_t^l\right). \tag{12}$$

3.5 Occlusion Handling

Although the appearance model is not updated during occlusions, these still constitute a main cause of catastrophic failures. Partial occlusions may cause inaccurate size updating, according to the area that can be seen. In case of complete occlusions, sample likelihoods are meaningless, and the re-sampling phase randomly propagate them, quickly losing the target.

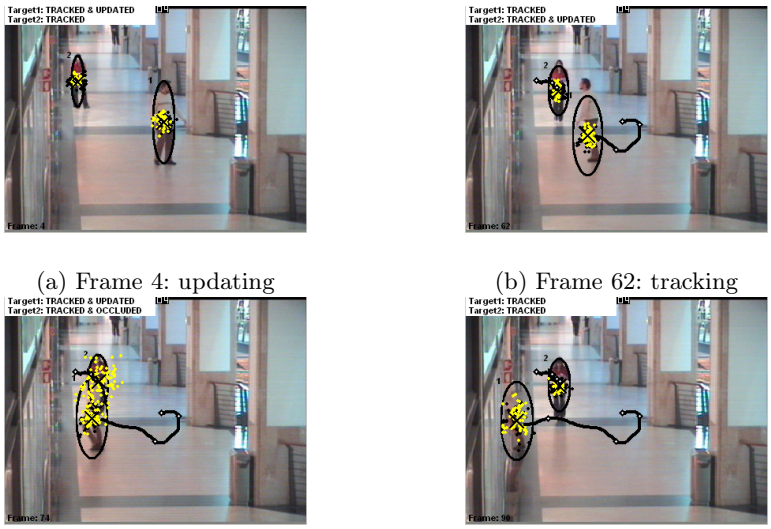
Hence, proper handling of occlusions is crucial. The state binary variable ρ_t^l tracks the occlusion status. Occlusions are predicted according to the learned dynamics. When the predicted occlusion is significant, and the target likelihood is lower than the expected one given by λ_t^l, the target state changes into occluded. Then, the following changes are introduced: (i) the adaptation rates are set to zero: neither the size, nor the velocity or the indicator itself is updated, and the position is just propagated; (ii) those samples belonging to the occluded target are not re-sampled. As a result, samples are spread around the target because of the uncertainty predictions terms. The other targets' samples are re-sampled, but are not assigned to the occluded target since otherwise this one would monopolise the whole sample set. When the occlusion is no longer predicted or a sample likelihood exceeds the value previous to the occlusion, ρ_t^l turns into 0, which immediately implies pruning those samples with lower weights. Furthermore, all estimates are again updated.

4 Experimental Results

The performance of the algorithm has been tested using the CAVIAR database[2]. In the sequence *OneLeaveShopReenter1cor* (Caviar database, 389 frames at 25 fps, 384 x 288 pixels), two targets are tracked simultaneously, despite their being articulated and elastic objects whose dynamics are highly non-linear, and that move through an environment which locally mimics the target colour appearance. The first target performs a rotation and heads towards the second one, eventually occluding it. The background colour distribution is so similar to the target ones that it constitutes a source of clutter. Furthermore, several oriented lighting sources are present, dramatically affecting the target appearance. Significant speed and size changes can also be observed.

The tracker performance is shown in Fig. 1. Both targets' appearance models are updated when reliable measures are obtained, see Fig. 1.(a). Poor localisations and occlusions are correctly detected, thereby avoiding re-sampling of samples of the occluded target and erroneous dynamic and appearance models updating, see Fig. 1.(b), (c). The tracker successfully recovers from occlusion, see

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR



(a) Frame 4: updating

(b) Frame 62: tracking

(c) Frame 74: occluded

(d) Frame 90: recovery

Fig. 1. Each target's estimated position is denoted by an ellipse and tagged accordingly; milestones are placed on the target trajectory every 25 frames; each predicted sample is drawn using a dark dot, whereas re-sampled particles are drawn in a light ones

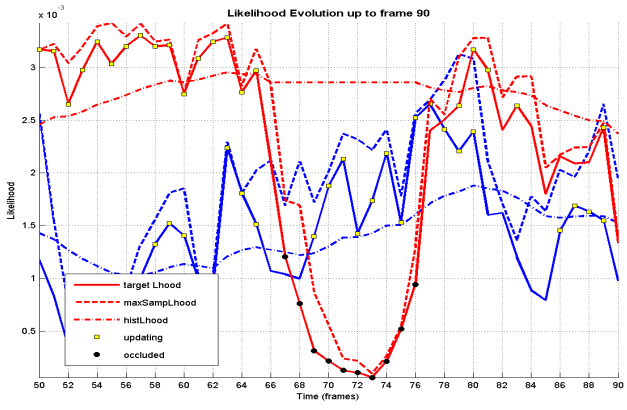


Fig. 2. Likelihood evolution

Fig. 1.(d). The maximum sample and target likelihoods, and the likelihood indicator is shown in Fig. 2. The tracker deals with multiple-target tracking whose dynamics are highly non-linear, despite using a simple constant speed approach. They move through an environment which mimics the target appearances. Furthermore, their trajectories intersect causing a severe partial occlusion. It copes with sizeable appearance and shape changes.

5 Conclusions

With this work we attempt to take a step towards solving the numerous difficulties which appear in unconstrained tracking applications. A robust likelihood function is used to properly evaluate samples associated to targets which present a high appearance variability. We rely on the Bhattacharyya coefficient between colour histograms to perform this task. Model updating is carried out with special care, thereby overcoming the model drift phenomenon. A multiple-target tracking scenario causes several problems, including sampling impoverishment and mutual occlusions. These issues are tackled by redefining the weight normalisation and predicting and handling occlusions.

The tracker has been successfully tested despite the fact that no detection is ever used after initialisation. Future research will be focused on careful feature selection in order to maximise the distance between the histograms corresponding to the different targets and the background, thereby enhancing the disambiguation of targets from clutter.

Acknowledgements. This work has been supported by the Catalan Research Agency (AGAUR), by the Spanish Ministry of Education (MEC) under projects TIC2003-08865 and DPI-2004-5414, and by the EC grant IST-027110 under the HERMES project.

References

1. Collins, R., Lipton, A., Kanade, T.: A System for Video Surveillance and Monitoring. In: 8th ITMRRS, Pittsburgh, USA, pp. 1–15. ANS (1999)
2. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based Object Tracking. PAMI 25(5), 564–577 (2003)
3. Deutscher, J., Reid, I.: Articulated Body Motion Capture by Stochastic Search. IJCV 61(2), 185–205 (2005)
4. Haritaoglu, I., Harwood, D., Davis, L.: W4: real-time surveillance of people and their activities. PAMI 22(8), 809–830 (2000)
5. Isard, M., MacCormick, J.: BraMBLe: A Bayesian Multiple-Blob Tracker. In: 8th ICCV, Vancouver, Canada, vol. 2, pp. 34–41. IEEE, Nagoya, Japan (2001)
6. Kahn, R., Swain, M., Prokopowicz, P., Firby, R.: Gesture Recognition Using the Perseus Architecture. In: CVPR, San Francisco, USA, pp. 734–741. IEEE, Nagoya, Japan (1996)
7. King, O., Forsyth, D.: How Does CONDENSATION Behave with a Finite Number of Samples? 6th ECCV, Ireland 1, 695–709 (2000)
8. MacCormick, J., Blake, A.: A Probabilistic Exclusion Principle for Tracking Multiple Objects. IJCV 39(1), 57–71 (2000)
9. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An Adaptive Color-Based Particle Filter. IVC 21(1), 99–110 (2003)
10. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based Probabilistic Tracking. 7th ECCV, Copenhaguen, Denmark. LNCS, pp. 661–675. Springer, Heidelberg (2002)
11. Rowe, D., Rius, I., González, J., Villanueva, J.J.: Improving Tracking by Handling Occlusions. 3rd ICAPR, UK. LNCS, vol. 2, pp. 384–393. Springer, Heidelberg (2005)