# A HJS FILTER TO TRACK VISUALLY INTERACTING TARGETS

*Oswald Lanz*

FBK-irst, Via Sommarive 18, 38100 Povo (TN), Italy

## ABSTRACT

Visual tracking with explicit occlusion models is computationally hard, in the sense that the complexity explodes as the number of targets increases. Recently, the Hybrid Joint-Separable (HJS) model has been proposed that enables tracking the local appearance of a number of bodies through occlusions with a quadratic, no more exponential, upper bound. In this paper we extend that method to account for a larger spectrum of visual interactions, captured by a full-image likelihood enabling true Bayesian inference, without compromising scalability. The resulting tracker then proves to be significantly more robust, and able to resolve long term occlusion among five people aligned on a single line-of-sight, observed from a single camera, at a manageable computational cost.

***Index Terms***— Visual tracking, Occlusion, Particle filter

## 1. INTRODUCTION

Visual inference in unconstraint scenes is always affected by uncertainty and ambiguity. This is of particular concern when tracking multiple bodies: uncertainty and ambiguity may derive from inaccurate interpretation of images, but can also be intrinsic in the measurement process, e.g. when occlusions exist, or in the monitored scene itself, e.g. when targets appear similar or clutter is present in the background. Bayesian methods allow to account for this in a principled way by representing estimates in form of distributions and relying on generative models of the observation process. Occlusions among tracked targets, which are recognized to represent a major source of failure for tracking systems, can then be modeled explicitly [1, 2]. When it comes to implementation, however, a computational problem is faced. Propagating the joint statistics of the different targets using generic representations is computationally hard, in the sense that the complexity becomes exponential in the number of bodies.

Many attempts have been made to find manageable solutions to multibody tracking. Partitioned sampling [2] avoids the high computational load associated with the joint approach by decomposing the joint state space into 1-body subspaces and performing updates separately and consecutively on them. In [3], only the Probability Hypothesis Density (PHD) of the multitarget posterior, i.e. its first moment, is propagated and its particle filter formulation becomes practical. In BraMBLe [1] estimation is accomplished jointly, in an expanded space that includes a discrete dimension reporting the number of tracked targets. Mixture tracking is proposed in [4], where each target is tracked as a single mode of a unique, multimodal distribution defined on a 1-body state space. A strategy for sampling in high dimensional spaces is proposed in [5], where a family of increasingly peaked likelihood functions are used to explore the state space gradually.

Our contribution here is over an approximate approach to multibody tracking [6]. The salient property of that approach is that it allows for tractable inference which is understood and theoretically grounded, and that it scales to input complexity, i.e. number of targets. A major limitation is a constraint on the likelihood function, which can account for the targets' local appearance only. As a consequence of that, the filter operates sub-optimally under a number of imaging conditions which occur frequently in practice. This work targets this limitation, and, to wipe it out, we show how to embed a more effective likelihood that accounts for a larger spectrum of visual interactions. While a straightforward upgrade of the standard HJS algorithm relapses it back to exponential complexity, we show how to maintain the inference tractable, preserving a quadratic upper bound, which is our key result.

Paper organization. Sec. 2 reviews the HJS framework, and reformulates the likelihood to accomodate our core contribution, which is presented and motivated in Sec. 3. Experiments are reported in Sec. 4, while Sec. 5 has the conclusions.

## 2. HJS VISUAL MULTIBODY TRACKING REVISED

Given a vectorial representation $\mathbf{x}$ of the state of a monitored environment in terms of object configurations it is composed of, the aim of Bayesian tracking is to estimate, at each time $t$, the posterior distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ conditioned on a sequence of observations $\mathbf{z}_{1:t}$ obtained up to $t$. This is done sequentially, by first propagating the posterior obtained at the previous time, $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$, according to a model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ of expected dynamics, and then updating it with the evidence contained in the new observation $z_t$ according to a model $l(z_t|\mathbf{x}_t)$ of the measurement process:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto l(z_t|\mathbf{x}_t)\int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\, d\mathbf{x}_{t-1}$$

Research partially funded by EU projects NETCARITY (PR 045508) and MY-E-DIRECTOR 2012 (PR 215248)

Since the complexity of computing this recursion increases exponentially with the size of $\mathbf{x}$, a fully joint formulation of the multibody problem becomes intractable even with a moderate number of targets. In particular, computing a reliable form of likelihood for a reasonably sized representation of the prior becomes prohibitively expensive. To allow for a tractable solution, the *Hybrid Joint-Separable (HJS)* model has been proposed. Instead of maintaining distributions in their joint form, a factorial representation in form of product of marginals $p(x_t^k|\mathbf{z}_{1:t}) = \int p(\mathbf{x}_t|\mathbf{z}_{1:t}) d\mathbf{x}_t^{\neg k}$ is estimated recursively ($x^k$ is the state of target $k$; $\mathbf{x}^{\neg k}$ is $\mathbf{x}$ with component $k$ removed). Although sub-optimal, this model enables for robust tracking through visual occlusions at an affordable computational cost. This is possible because updating the marginals does not require, per se, explicit computations in the joint domain. To show this we introduce a likelihood that accounts for (i) visual occlusions, and (ii) statistical independence of object blobs (details are in Sec. 3):

$$-\log l(z|\mathbf{x}) = \sum_k L^k(z|\omega_{\mathbf{x}}^k), \tag{1}$$

where $[\omega_{\mathbf{x}}^k]_u$ represents a binary mask assuming values 1 in pixels $u$ where target $k$ is the closest to the camera, while $L^k$ describes some generic form of visual dissimilarity of the portion of $z$ with support $\omega_{\mathbf{x}}^k$ and a reference model of target $k$, e.g. a distance between color histograms. If $\mathbf{x}$ conveys information about the targets' distance to the camera, then $\omega_{\mathbf{x}}^k$, which encodes data association under $\mathbf{x}$, can be computed using shape projections. Therefore, we (i) describe the state of each target by its position on a horizontal reference plane (e.g. the floor), (ii) adopt a generalized-cylinder model for target shape, and (iii) assume calibrated camera.

It has been shown [6] that, with the range of the likelihood in Eq. 1 restrained to the target's local appearance, i.e. to the image region confined by shape model projection, a reliable approximation to the marginals can be calculated with a quadratic cost. This is done by updating each factor $p(x_t^k|\mathbf{z}_{1:t-1})$ of the separable temporal prior with a marginalized log-likelihood

$$L^k(z_t|w_{x_t^k}^k) + \sum_{l\neq k} \int L^l(z_t|w_{\mathbf{x}_t^{kl}}^l) p(x_t^l|\mathbf{z}_{1:t-1})\, dx_t^l \tag{2}$$

where $w_{\mathbf{y}}^j$ is zero everywhere but on pixels $u$ that are internal to the projected shape of $x_t^k$, there taking values

$$[w_{\mathbf{y}}^j]_u = 1 - \prod_{i\neq \mathrm{idx}(\mathbf{y})} \int [\omega_{x^i}^i]_u\, p(x^i|\mathbf{z}_{1:t-1})\, dx^i. \tag{3}$$

Here $w_{\mathbf{y}}^j$, with $\mathbf{y} \in \{x_t^j, \mathbf{x}_t^{jk}\}$ and $\mathrm{idx}(\mathbf{y}) \in \{j, jk\}$, is now a smooth image kernel providing pixel weights to be accounted for by $L$. The advantage is that marginalization is transferred to image masks $\omega_{\mathbf{x}}^k$, where it can be partitioned, thus computed efficiently. By doing so, the more expensive evaluation of $L^j$ is done at most $K$ times for each state ($K$ is the number of tracked bodies). We will refer to Eq. 2 as the HJS log-likelihood of target $k$. A detailed mathematical derivation is available in [6], which holds under locality.

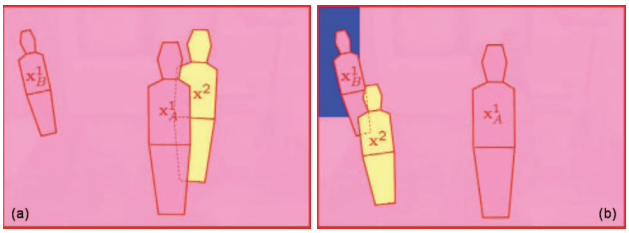


**Fig. 1**. Two examples which highlight the limitations of likelihoods whose influence is confined to the target's own support, like [6]. In both cases exact localization is possible only with the proposed extension. See text, and experiments in Sec. 4.1.

## 3. HJS LOG-LIKELIHOOD FOR VISUALLY INTERACTING TARGETS

To motivate our contribution we characterize imaging conditions under which the standard HJS filter operates sub-optimally, with excessive uncertainty. Two core examples are depicted in Fig. 1. According to [6], dissimilarity terms $L^l$ in Eq. 2 are computed over the image area delimited by shape projection. Consequently, the marginalized likelihoods for $x_A^1$ and $x_B^1$ return the same (maximal) value, and this holds for both figures. In the context of multibody tracking this can lead to artifacts such as maintenance of *phantom modes* in the posterior after an occlusion has occurred. A tracker that discards non-local information (i.e. information about where the target cannot be) is here prone to fail if the noise in the appearance of the previously occluded target is higher than the one of the occluder. To avoid this, we build on the following observation: *$x_A^1$ is the true state of target 1 in Fig. 1(a) because it uniquely explains the occlusion on $x^2$.* Thus, $L^1$ must be reformulated to get influenced by the $L^2$ value over $x^2$, in a way that $x_B^1$ gets penalized by the hypothesized unoccluded appearance of $x^2$. To infer $x_B^1$ as the unique plausible state for Fig. 1(b), that influence must be extended even more, to the background. These modifications introduce a strong, non-local, visual interaction.

To account for non-local information, we now design a likelihood whose influence is no longer restricted to the target's shape projection, but extends to a function of the entire image. Given a state $\mathbf{x}$, the image is tessellated into object blobs and background through shape projections (Fig. 1). Occlusions are handled by associating a pixel to the object that is closer to the camera along the considered line-of-sight. The remaining part is decomposed into a grid of $N_x \times N_y$ regular cells $\omega_{i,j}^0$. All patches are supposed to have mutually independent appearance, so $l(z|\mathbf{x})$ factorizes over them and Eq. 1

is obtained. Since we now want to import this model as-it-is, i.e. without restricting its range to the local support of the target considered, we can no longer apply the original algorithm. Likelihood marginalization would have to be done explicitly, thus relapsing the method back to exponential complexity.

An effective embedding is possible after revising Eq. 2. The first term, $L^k$, remains unchanged because it is computed using an image kernel which, by its definition, has local support. The terms under the integrals, $L^l$, now use a kernel that expands beyond the image area delimited by the projected shape of $x_t^k$. The external part of it, say $m^l$, still takes his non-zero values according to Eq. 3. The complete kernel is thus assembled pixel-wise, by $[w^l]_u + [m^l]_u$, or $w^l \oplus m^l$ in compact notation. Under the assumption that the appearance error spreads uniformly over the support of target $l$, i.e. $L^l(z|w^l \oplus m^l) \approx L^l(z|w^l) + L^l(z|m^l)$ we can partition the computations: first the standard method is applied to compute the local contribution (Eq. 2), which is then upgraded with the term

$$\sum_{l \neq k} \int L^l(z_t|m^l_{\mathbf{x}_t^{kl}}) p(x_t^l|\mathbf{z}_{1:t-1})\, dx_t^l. \tag{4}$$

The HJS likelihood now accounts for visual interactions among all targets, being they occluding each other or not. To seamlessly integrate a tessellated model of the background (needed, e.g., to solve Fig 1(b)) we assign a fixed, deterministic posterior to each patch $\omega_{ij}^0$. Eq. 2 and Eq. 4 then apply as before, with the summation including the terms for $l = 0$. While the introduction of a full-image likelihood now enables for true Bayesian inference, in some applications reliable background information may not be available or difficult to maintain. In such cases the extended filter can still be applied to systematically handle the ambiguities in Fig 1(a).

The standard algorithm (HJS particle filter, see [6]) is first applied to compute HJS likelihoods on the local support. A pair of image buffers per target are used to this purpose: each $B^l_{\text{ker}}$ contains one factor of Eq. 3, while from $B^l_{\text{evd}}$ the sum terms in Eq. 2 are extracted. After this first step an occlusion kernel, $O^k$, is assembled for each target $k$ from the set $\{B^l_{\text{ker}}\}_{l \neq k}$ following Eq. 3. Each particle is then revisited, and the HJS likelihood is updated as follows. The background term ($l = 0$) of Eq. 4 is computed using the occlusion kernel with the shape projection area inhibited, i.e. with the values internal to the projection set to zero. The remaining terms of Eq. 4, which account for extended interactions with the other targets, are integrated by summing up all $B^l_{\text{evd}}$ values that are external to the projected shape. To speed up, the demanding computation of the background term can be skipped as follows: instead of scoring the background for each single particle using its shape projection as kernel we can do it once for all, by including $B^l_{\text{ker}}$ in the calculation of the occlusion kernel and updating $B^l_{\text{evd}}$ correspondingly, before the second pass. This way the binary mask is replaced with its expectation, thus occlusion boundaries are smoothed out, resulting in a less peaked likelihood, which is the price to pay for speedup.

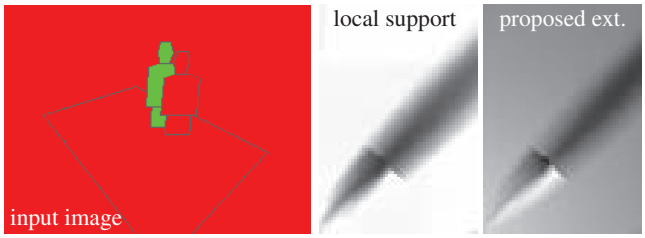


**Fig. 2**. Input image A (with target outlines and domain boundary overlayed in gray), and HJS likelihood response for the red target with local support only ([6], center) and proposed extension (Sec. 3, without background term, right image).

The overhead introduced is linear in the overall representation size, which is number of targets $K$ times number of particles $N$ used to represent each factor of the posterior. The algorithm's scalability is therefore left unchanged, which is our key result: computational complexity remains $O(K^2N)$.

## 4. EXPERIMENTS

### 4.1. Likelihood response on synthetic data

To verify the claims made in Sec. 3 which led to the proposed extension we compare original and proposed method on two synthetic examples. Fig. 2 shows an image with a green figure partially occluded by a red figure over a red background. To the right there are the gray coded likelihood values computed over a uniform grid placed on the floor for the red figure, using the two methods. The position of the green figure is known. With the locally defined likelihood it is not possible to localize the target due to the cluttered background. Indeed, the likelihood peek spans uniformly over a large portion of the domain. The proposed extension, which accounts for additional evidence on the support of the green figure, suppresses hypotheses that map onto the background because they do not explain the partial occlusion on the green figure. This results in a peaked response, from which the red figure can be localized precisely. Fig. 3 shows the figures drawn with a regular color pattern (each other pixel has lower intensity), simulating a noisy observation. The red figure has more contrast in the pattern, thus its color model is noisier than that of the green figure. The occlusion robust formulation of the HJS likelihood makes the original version priviledge occluded hypotheses because of the less noisy color model of the occluder. A particle filter using such likelihood would, after an occlusion, maintain the higher scored particles in the occlusion volume of the green target, generating a *phantom mode* in the estimates, and potentially loosing its track after resampling. With the extension that mode is again suppressed. This makes the tracking significantly more robust in situations with unevenly noisy target models, which are often captured in real world applications during online operation.

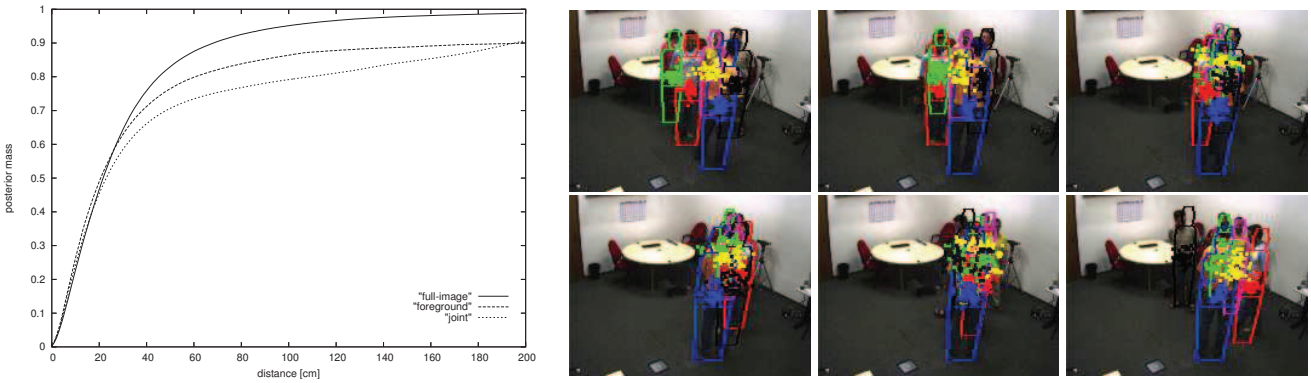


**Fig. 4**. Localization error distribution on test sequence, and key frames. See http://tev.fbk.eu/smartrack/icassp09.avi for a video.

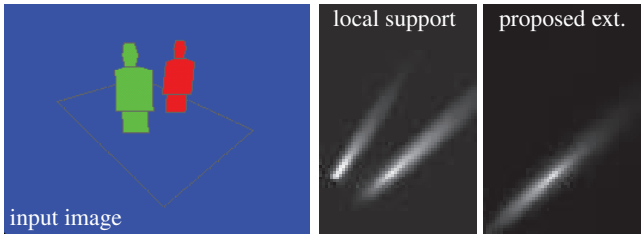


**Fig. 3**. Likelihoods on image B (including background term).

### 4.2. Tracking on real data

To verify expected improvements we track five people through a monocular sequence (50s) of scaled difficulty. In the first 20s the dynamics is high, with many partial and complete occlusions occurring between up to three targets. After 25s all five people arrange along a line-of-sight for about 5 secs ($> 70$ frames). During this period only the person in front is visible, with all the others almost completely occluded.

Appearance models are acquired manually, Bhattacharyya coefficient is used for likelihood evaluation, and 150 particles are assigned to each target. All tracks are initialized manually. Image resolution is $200 \times 150$. Processing the whole sequence with the standard filter took 10.7 secs, and 42.9 secs with the full extension, including background term. The first method diverges in the initial phase of complete occlusion, and cannot recover. The enhanced version of the same filter keeps track of all targets across the whole sequence (Fig. 4).

To give quantitative results we report in Fig. 4 the amount of posterior mass cummulated around the ground truth as a function of distance from the ground truth position (i.e. the localization error). The improvement achieved is significant, and underlines the necessity of introducing non-local interactions. To highlight efficiency, we compare it with a fully joint implementation of the same tracker, which would correspond to a straightforward extension of the work in [6]. $10^5$ joint particles were used, with each frame taking about 6.5 secs to be processed. In spite of excessive processing load the filter failed soon because of sparse sampling of the joint space. The presented contribution, together with the HJS framework, allows to keep the workload manageable, while maintaining robustness and accuracy at a comparable level.

## 5. CONCLUSION

In the context of multibody tracking we have (i) characterized the limitations of a class of visual likelihoods that account for the target's local appearance only, and (ii) proposed a HJS filter build upon a revised likelihood that now captures a large spectrum of visual interactions. Regardless of the 'curse of dimensionality' inherent in the problem, the resulting algorithm still scales to input complexity, i.e. number of targets, with quadratic upper bound. While improvements in terms of robustness and accuracy were demonstrated on a challenging sequence, we expect that this contribution will reveal its real value in the context of articulated motion tracking.

## 6. REFERENCES

[1] M. Isard and J. MacCormick, "BraMBLe: A bayesian multiple-blob tracker," in *ICCV*, 2003.

[2] J. MacCormick and A. Blake, "Probabilistic exclusion and partitioned sampling for tracking multiple objects," *IJCV*, vol. 39, no. 1, 2000.

[3] B. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for Bayesian multi-target filtering with Random Finite Sets," *IEEE Trans. AES*, vol. 41, no. 4, 2005.

[4] J. Vermaak, A. Doucet, and P. Perez, "Maintaining multi-modality through mixture tracking," in *ICCV*, 2001.

[5] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *CVPR*.

[6] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Trans. PAMI*, vol. 28, no. 9, 2006.