A Comparative View on Exemplar 'Tracking-by-Detection' Approaches

Elie Moussy, Alhayat Ali Mekonnen, Guilhem Marion, Frédéric Lerasle

A Comparative View on Exemplar 'Tracking-by-Detection' Approaches *

E. Moussy[1], A. A. Mekonnen[1], G. Marion[1,2], F. Lerasle[1,2]
[1]CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France
[2]Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
{emoussy,aamekonn,marion,lerasle}@laas.fr

Abstract

In this work, we present a comparative evaluation of various 'tracking-by-detection' approaches on public datasets. The work investigates popular sequential Monte Carlo and template ensemble based trackers coupled with relevant visual people detectors with emphasis on exhibited performance variation depending on tracker-detector choice. Extensive experimental results are provided on public dataset and results indicate the choice of a detector can significantly vary the performance of a tracker. Our experimental results show, depending on the choice of the detector, average tracking accuracy across three public datasets could exhibit a 45% *standard deviation with only, on average, a* 6.8% *and* 11.1% *standard deviation in detector recall and precision respectively.*

1. Introduction

People detection and tracking is an important research area with prominent applications in video surveillance, pedestrian protection systems, human-computer interaction, robotics, and the like. As a result, it has amassed huge interest from the scientific community [2, 5, 4]. People tracking falls under Multi-Object Tracking (MOT) which deals with the process of accurately estimating the state of objects – position, identity, and configuration – over time from observations. Due to incurred challenges – scene clutter, target dynamics, intra/inter-class variation, measurement noise, frame rate – it has long been established that coupling trackers with detectors, in a paradigm called 'tracking-by-detection', helps better tackle these challenges [2, 9, 8]. In the context of people tracking, 'tracking-by-detection' approaches rely on a people detector to start, update, re-initialize, guide (avoid drift), or terminate a tracker.

*This work was supported by grants from the French General Directorate for Armament (DGA) under grant reference SERVAT RAPID-142906073 and the French National Research Agency (ANR) under project RIDDLE with grant number ANR-12-CORD-0003.

In the literature, it is common to find many 'tracking-by-detection' approaches applied to people tracking. However the usual trend it to select a single detector and directly couple it with the tracker, e.g, [2, 9]. With the advent of various different people detection techniques with significant variations in terms of detection performance and speed, see [5], the first step should be to evaluate the effect a detector choice impacts performance, and the relevant association with filtering strategies. To the best of our knowledge no such work exits till date. To clarify, there are indeed very good experimental comparative works in detection, e.g., [5], but none that shows the inter related effects by detector and tracker choices. To bridge this gap, in this work, we present a comparative evaluation of exemplar 'tracking-by-detection' approaches, with different detectors, on relevant public datasets, primarily focusing on sequential Monte Carlo approach as most approaches rely on it. We consider three different trackers (filtering strategies) owing to their pervasive use in the literature and relevance: A Decentralized Particle Filter (DPF), e.g., [2, 7], Tracker Hierarchy [12], and Reversible Jump Markov Chain Monte Carlo - Particle Filter (RJMCMC) [8]. The DPF and RJMCMC are selected as they are the most popular Monte Carlo approaches marking two important tracker configurations: a decentralized one which assigns an independent tracker per target, and centralized one – also called a joint state tracker – in which all the states of the tracked targets are concatenated forming a single representation that captures the entire configuration. The Tracker Hierarchy is selected as it, unlike the usual implementations of DPF and RJMCMC which utilize simple single target representation, consists of a rich target representation model in the form of template ensembles. Similar to DPF, it is a decentralized approach and has shown tracking results comparable to the state-of-the-art [12].

The above mentioned trackers are coupled with three selected detector, namely: Dalal and Triggs Histogram of Oriented Gradients (HOG) based detector [3] denoted as HOG-SVM, Felzenszwalb et al. [6] Deformable Part-based Methods (DPM) detector, and Dollar et al. [4] Aggregate

Channel Features (ACF) based detector. The detectors mark three distinct detector 'superiority era' onsets as published in 2005, 2010, and 2014 consecutively. Furthermore, our choice is motivated by the fact that both ACF and DPM are amongst the current best detectors and HOG-SVM, though not currently the best itself, its features make constituents, one way or another, of current state-of-art approaches and historically has been the *de facto* benchmark detector.

In short, the selected trackers and detectors are quite relevant in the literature and the variations amongst them, namely: (1) tracker configuration that governs how the state space is explored, (2) variation in target representation – simple color appearance to template ensemble, (3) the detectors variation in terms of detection accuracy and precision [5], dictates the representativeness and relevance of this comparative evaluation. Hence, the contributions of this paper are: (1) systematic evaluation of 'tracking-by-detection' approach based on relevant combination of trackers and detectors, (2) a presentation of relevant results with insightful discussions that highlight the benefits of tracker choice, target representation, and detector choice.

This paper is organized as follows: section 2 starts with background and overview of the different trackers and detectors combined, section 3 details the experimental setting and obtained results, section 4 presents a comprehensive discussion, and the paper finally finishes off with conclusions and future works in section 5.

2. Background and overview

In this section, we will provide an overview and relevant background information about the two main ingredients of 'people tracking-by-detection': people detection and tracking. The different detectors and trackers selected for the comparative evaluation are briefly described.

2.1. People Detection

Depending on the type of feature, model, classifier, and learning technique adopted, visual people detectors perform variably on different datasets. Below, the three chosen detectors are briefly explained.

HOG-SVM [3]: This detector is one of the classical and oldest detectors. This detector computes local histograms of the gradient orientation on a dense grid and uses linear Support Vector Machine (SVM) as a classifier. The constituent HOG features have shown to be the most discriminant features to date, and in fact, a majority of detectors proposed hence-after make use of HOG or its variant one way or another [5].

DPM [6]: Contrary to HOG and ACF which detect a person's full body, the DPM is a parts based detector that works by aggregating evidence of different parts of a body to detect a person in an image. Its trained model is divided

(a) HOG-SVM (b) DPM (c) ACF

Figure 1: Sample detection outputs of the three detectors.

in different parts. For instance, a person's model could be made up of a head, upper body, and lower body sub-models. Each detected area has its own score. Thus, it is possible to put a threshold in order to remove detections that have low scores. Since this detector relies on parts, and not solely on a full body, it detects partially occluded people rather well, see figure 1b for example. Additionally, it also has better localization accuracy as it infers bounding box based on detected body parts. The detector uses variants of HOG features with Latent-SVM as classifier.

ACF [4]: This is a fast person detector that has shown state-of-the-art performance on various benchmarking datasets [4]. It is based on aggregates (summed over blocks) of features represented as channels and a variant of Boosted classifier. Examples of features channels used include: normalized amplitude of the gradient, the histograms of oriented gradients (HOG, 6 channels) and color channels (LUV).

Generally speaking, the ACF, amongst the others, does significantly better in outdoor environments, e.g., camera mounted on a vehicle, whereas the DPM outperforms whenever the dataset contains partially occluded person instances. Surprising, the HOG-SVM detector also does well with the presence of contrasting background that clearly helps delimit the boundaries of people. As a result, it is the interest in this work to show how these detector performance variations affect 'tracking-by-detection' based trackers, notably their filtering and target appearance model robustness, on different benchmarking public datasets. Sample detections are shown in figure 1.

2.2. Multi-Object Tracking (MOT)

For MOT, we consider the two popular tracker configurations: A decentralized and centralized tracker configuration.

2.2.1 Decentralized MOT

In decentralized MOT, each instantiated tracker has its own state vector which is independent of the others. In this class, we chose to use our variant of the classical Particle Filter (PF) – a popular choice in the literature, e.g., [2, 7] – and the more involved open-sourced Tracker Hierarchy [12] which is a template ensemble based tracker.

Decentralized Particle Filters (DPF): In this approach, each target is assigned a unique instance of a Particle Filter as a tracker. In this work, this target specific tracker is

implemented based on the ICondensation [10] filter as it is the most widely used and suitable PF variant for detector integration. This is a sequential Monte Carlo approach which approximates the posterior over the target state x_t given all measurements up to time t, $Z_{1:t}$, using a set of N weighted samples, i.e., $p(x_t|Z_{1:t}) \approx \{x_t^{(i)}, w_t^{(i)}\}_{i=1}^{N}$. Tracking is achieved sequentially with the notion of Importance Sampling whereby the particles at time $t-1$ are propagated according to a proposal density, $q(.)$, and their weights are updated in accordance with equation 1.

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(z_t|x_t^{(i)})p(x_t^{(i)}|p(x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)}, z_t)} \tag{1}$$

Where, $p(x_t^{(i)}|x_{t-1}^{(i)})$ is target dynamic model, $p(z_t|x_t^{(i)})$ is likelihood term, and $q(x_t^{(i)}|x_{t-1}^{(i)}, z_t)$ is the proposal density evaluated at the sampled state. To derive this filter with incoming detections, hence to realize 'tracking-by-detection', the proposal density shown in equation 2 is employed. According to this density, part of the particles will be sampled from the detector cues, $\pi(x_t^{(i)}|z_t)$, some from the dynamics, and some from the prior $p_0(x_t^{(i)})$, in accordance with the ratios α, β, γ which should sum to 1.

$$q(x_t^{(i)}|x_{t-1}^{(i)}, z_t) = \beta p(x_t^{(i)}|x_{t-1}^{(i)}) + \alpha\pi(x_t^{(i)}|z_t) + \gamma p_0(x_t^{(i)}) \tag{2}$$

The likelihood $p(z_t|x_t^{(i)})$ is a probabilistic measure based on the Bhattacharyya similarity coefficient, as in [10], with respect to a dynamically updated simple single target color histogram, constructed in the HSV color space. During MOT, each incoming detections have to be associated with the different trackers distinctly. For that we use a greedy assignment algorithm [2] which performs comparably to the famous Hungarian assignment algorithm. The DPF is coupled with each one of the detectors presented in section 2.1 during experimental evaluations. The exact label for the tracker is derived by appending the detector name on DPF. For example, DPF-ACF is the decentralized Particle Filter multi-object tracker using ACF as detector.

Tracker Hierarchy (Hierarchy) [12]: This multi-object tracker is another 'tracking-by-detection' decentralized MOT that assigns a single tracker per target. It is a tracker that consists of a rich appearance model of the target in the form of a template ensemble and uses hierarchy of expert and novice trackers for efficient multi-person tracking. At each time step, the correct target position is estimated by making use of a mean-shift mode estimator and a Kalman filter. We consider evaluating its performance with different detectors as it, unlike the two other trackers, consists of a rich and involved target appearance model. Please refer to [12] for further details. This tracker combined with any of the detectors is labelled as Hierarchy followed by detector name, e.g., Hierarchy-ACF.

2.2.2 Centralized MOT

The centralized MOT, also called joint state tracker, represents the tracked target using a joint state that captures the entire configuration. The main advantage is that should the targets interact, an interaction model can be incorporated in the tracking step. The most famous approach in this class is the **Reversible Jump Markov Chain Monte Carlo - Particle Filter (RJMCMC)** tracker [8]. The RJMCMC defines a Markov Chain over the state configuration so that the stationary distribution of the chain approximates the posterior distribution $p(x_t|Z_{1:t})$. It replaces the inefficient important sampling step with a trans-dimensional Metropolis Hastings (MH) algorithm to sample from the chain.

In RJMCMC tracker, the posterior at time t is approximated using a set of M discrete unweighted samples: $p(x_t|Z_{1:t}) \approx \{x_{t-1}^{(i)}\}_{i=1}^{M}$. It uses a set of moves m to change the dimension of the state, i.e., adding new target, removing untracked targets, or leave it unchanged according to a prior move proposal q_m. Each move is associated with a move specific proposal distribution $Q_m(.)$. Each move m must have a reverse move $m*$ that assures reversibility so that detailed balance will be achieved and the chain will converge to the desired stationary distribution [8]. During the iterative estimation process, at the i^{th} iteration, it first samples a move from q_m and proposes a new particle x^* based on $Q_m(.)$. It then computes the acceptance ratio α_a shown in equation 3 considering $Q_m(.)$ and the reverse move proposal distribution $Q_{m^*}(.)$. The proposed particle is accepted with probability α_a or otherwise rejected. Particles used both for the *burn-in*, M_b, and *thin-out*, M_{th}, are discarded leaving M unweighted samples to represent the posterior.

$$\alpha_a = \min\left(1, \frac{p(x^*|Z_{1:t})Q_{m^*}(x_t^{(i-1)}; x^*)q_{m^*}\Psi(x^*)}{p(x_t^{(i-1)}|Z_{1:t})Q_m(x^*; x_t^{(i-1)})q_m\Psi(x_t^{(i-1)})}\right) \tag{3}$$

$\Psi(.)$ in equation 3 is the interaction model. Our implementation is based on [8] with the set of moves $m = \{add, delete, stay, leave, update, swap\}$, a markov random field based interaction model $\Psi(.)$, a likelihood measure $p(z_t|x_t)$ based on the Bhattacharyya similarity coefficient of a simple single target color histogram, contrary to Hierarchy which has an ensemble, in the HSV color space. The different move specific proposal distributions, $Q_m(.)$, are defined as in [8]. The RJMCMC is coupled with the different detectors presented and evaluated. The coupled tracker-detector is denoted using RJMCMC followed by used detector acronym, e.g., RJMCMC-ACF. Similar to DPF, the detection-track data association is handled via a greedy assignment algorithm.

3. Experiments and results

As stated, the main objective of this paper is to provide a comparative evaluation of three multi-object 'tracking-by-detection' based tracking techniques utilizing three different detectors. The tracker and detector combination leads to nine evaluated combinations. In this section, we present the different datasets, evaluation metrics, implementation details, and obtained results.

3.1. Datasets

To evaluate the different detector-tracker combination, we have utilized three public datasets, namely: The CAVIAR OneShopOneWait dataset[1], the CAVIAR EnterExitCrossingPaths dataset[1], and the PETS2009S2L1[2] dataset. Here onwards, these datasets are referred as *OneShop*, *EnterExit*, and *PETSS2L1* respectively. These datasets are selected as they highlight different conditions: (1) the OneShop sequence consists of intermittent target occlusions, (2) the EnterExit sequence is challenging for detectors due to background clutter, and (3) the PETSS2L1 features an outdoor scene with many targets and encountered target-target interactions.

Sequence	Frame Rate	No. Frames	No. of Id.	Detector Recall / Precision		
				ACF	DPM	HOG-SVM
EnterExit	25	383	4	.58 / .76	**.74** / **.89**	.62 / .77
OneShop	25	1377	6	.32 / .48	**.63** / **.90**	.44 / .43
PETSS2L1	7	795	19	**.88** / **.93**	.80 / .92	.80 / .90

Table 1: Utilized public datasets along with detector performance.

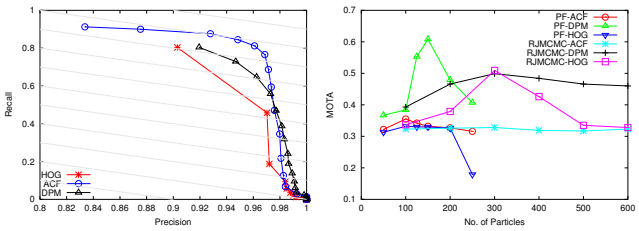
Relevant descriptions of the different datasets along with the performance of the three detectors is provided in table 1. At this stage, the detector's performance is quantized as recall = $\frac{TP}{TP+FN}$ and precision = $\frac{TP}{TP+FP}$, where TP is for true positives, FP for false positives, and FN for false negatives.

3.2. Evaluation metrics

To quantify the performance of the different trackers, we utilize the prevalent CLEAR-MOT metrics [1]. The CLEAR-MOT metrics are principally based on computation of two quantities: the Multi-Object Tracking Accuracy (MOTA) and the Multi-Object Tracking Precision (MOTP). The MOTA = $1-(\mathcal{F}_P+\mathcal{F}_N+\mathcal{I}d_{sw})$, where $\mathcal{F}_P = \sum_t \frac{FP_t}{g_t}$ quantifies total false positives, $\mathcal{F}_N = \sum_t \frac{FN_t}{g_t}$ quantifies the total false negatives, and $\mathcal{I}d_{sw} = \sum_t \frac{Id_{sw,t}}{g_t}$ quantifies total id switches, all divided by the ground truth targets and summed over the entire dataset. The MOTP is the average bounding box overlap between the estimated target position and ground truth annotations over the correctly tracked targets. A tracker estimated rectangular position, R_T, is considered a correct track if its overlapping area score, $sc = \frac{R_T \cap G_T}{R_T \cup G_T}$, with the ground truth annotation G_T is above a threshold τ, which is usually set to 0.5.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[2] http://www.cvg.reading.ac.uk/PETS2009/a.html



(a) (b)

Figure 2: (a) Detector Precision-Recall curve on a subset of the PETSS2L1 dataset. (b) Performance, in terms of MOTA, of the Monte Carlo trackers as a function of number of particles used to approximate the posterior.

3.3. Implementation details

Detailing implementation specific choices, for the detectors: for HOG-SVM, we use GPU implementation of OpenCV library[3] with a model trained on the INRIA dataset, for DPM and ACF, we use the original Matlab source codes discussed in [6] and [4] respectively. To further tune the threshold of the detectors, we evaluate them on a subset of the PETSS2L1 dataset with the help of Precision-Recall curves. Accordingly, as shown in figure 2a the operating point is set at the extreme point where the curve touches the faded gray lines (the lines denote equal trade-off between precision-recall). For DPM and HOG-SVM, this is the point where the highest recall is attained, whereas for ACF it is where it achieves a 0.93 precision. All the results in table 1 are obtained at these operating points.

With regards to the trackers, for Hierarchy, we use the original source code [12] with only the dataset frame rate parameter adjusted accordingly. Both the DPF and RJMCMC are implemented in C++. To validate the number of particles to use in each tracker, they are evaluated using a subset of the EnterExit dataset by varying the number of particles used to approximate the posterior, shown in figure 2b. As the result shows, the best performance averaged over the different detector is obtained when using $N = 150$ particles for DPF, and $M = 300$ particles for RJMCMC. Furthermore, for RJMCMC, the *burn-in* and *thin-in* particles are set to $M_b = 45$ and $M_{th} = 3$ respectively, and $q_m = \{0.1, 0.01, 0.01, 0.05, 0.82, 0.01\}$. For DPF, proposal density weights of $\beta = 0.6$, $\alpha = 0.4$, $\gamma = 0$ are used. These are also exactly the same for RJMCMC during the update move. In both RJMCMC and DPF, a random walk dynamic model is utilized. We do not report on computational speed of the trackers as we are using unoptimized code that has a mix of Matlab (for two of the detectors) and C++ code.

[3] http://opencv.org/

3.4. Results

The results obtained by running the tracker-detector combinations on the three public datasets are shown in detail in tables 2, 3, and 4, corresponding to the EnterExit, OneShop, and PETSS2L1 datasets respectively. Each tracker-detector combination is run five times to account for their stochastic nature. In all cases, the results are reported as mean/standard deviation, i.e., μ/σ. Before analyzing these results, it is worth taking a look at the detector performance on these datasets in table 1. All the three detectors achieve high detection rates, $> 80\%$, with high precision, $> 90\%$, on the PETSSL1 dataset with top performance from ACF detecting 88% percent of the targets correctly. On the contrary, ACF does poorly on the other two datasets, EnterExit and OneShop, with DPM achieving the best result trailed by HOG-SVM. On average, DPM perform much better than the others with consistent high precision rates. The results also indicate that the OneShop dataset poses a great challenge for detectors followed by EnterExit.

Tracker-Detector	MOTA	MOTP	$\mathcal{T}_P$	$\mathcal{F}_P$	$\mathcal{F}_N$	Id_{sw}
DPF-ACF	.25 / .02	.42 / .04	.77 / .09	.52 / .02	**.21** / .07	19.4 / 10.2
Hierarchy-ACF	**.31** / .00	**.64** / .00	**.68** / .00	.38 / .00	.31 / .00	**6.9** / 0.0
RJMCMC-ACF	.30 / .02	**.64** / .02	.55 / .02	**.24** / .01	.43 / .02	32.8 / 6.7
DPF-DPM	.68 / .01	.59 / .04	.79 / .07	.12 / .06	.19 / .07	**1.5** / 0.4
Hierarchy-DPM	**.75** / .00	**.76** / .00	**.82** / .00	.08 / .00	**.17** / .00	4.2 / 0.0
RJMCMC-DPM	.53 / .02	.62 / .01	.54 / .03	**.01** / .01	.45 / .03	9.0 / 2.7
DPF-HOG	.23 / .07	.62 / .01	.60 / .04	.37 / .05	.41 / .02	**1.9** / 1.2
Hierarchy-HOG	.24 / .00	.22 / .00	.36 / .00	.12 / .00	.63 / .00	18.0 / 0.0
RJMCMC-HOG	**.50** / .04	**.41** / .00	**.58** / .05	.08 / .04	**.40** / .05	6.9 / 5.7

Table 2: CLEAR-MOT results on the EnterExit dataset. All results reported as μ/σ computed over five runs.

Table 2 shows the detailed results on the EnterExit dataset. The best overall result, both accuracy and precision, is obtained with the Hierarchy-DPM tracker. The best results of each of the trackers is obtained when combined with the DPM detector. As the ACF performs poorly in this dataset, corresponding results are accordingly poor. In this dataset, the DPF is best combined with ACF, the Hierarchy with DPM, and the RJMCMC also with DPM. The HOG-SVM detector is better combined with the RJMCMC tracker.

Tracker-Detector	MOTA	MOTP	$\mathcal{T}_P$	$\mathcal{F}_P$	$\mathcal{F}_N$	Id_{sw}
DPF-ACF	-.22 / .07	.43 / .03	.48 / .03	.71 / .07	.50 / .03	**9.9** / 5.9
Hierarchy-ACF	-.50 / .00	.51 / .00	**.55** / .00	1.05 / .00	**.42** / .00	121.3 / 0.0
RJMCMC-ACF	**.00** / .01	**.64** / .02	.10 / .01	**.10** / .01	.88 / .01	75.4 / 6.2
DPF-DPM	**.58** / .32	.59 / .02	**.69** / .05	.10 / .07	**.29** / .04	**16.2** / 5.1
Hierarchy-DPM	.39 / .00	**.68** / .00	.58 / .00	.19 / .00	.38 / .00	184.3 / 0.0
RJMCMC-DPM	.41 / .00	.59 / .00	.41 / .01	**.01** / .01	.58 / .01	27.4 / 5.4
DPF-HOG	.09 / .08	.26 / .01	**.54** / .03	.45 / .06	**.45** / .04	**8.1** / 4.9
Hierarchy-HOG	**.29** / .00	.33 / .00	.33 / .00	**.04** / .00	.67 / .00	26.9 / 0.0
RJMCMC-HOG	.16 / .02	**.42** / .00	.31 / .02	.15 / .03	.69 / .02	20.8 / 5.5

Table 3: CLEAR-MOT results on the OneShop dataset. All results reported as μ/σ computed over five runs.

Table 3 shows the results on the OneShop dataset. This is the dataset where the detectors all perform poorly. This is clearly reflected in the results. Again the best results, precision and accuracy, involve the DPM detector: accuracy with DPF-DPM and precision with Hierarchy-DPM. The best result, a 58% MOTA, is achieved with the DPF-DPM tracker. Besides, when the detector is not reliable such as the ACF in Table 3, the DPF and the RJMCMC handle it better than the Hierarchy, i.e., have better filtering capabilities.

Tracker-Detector	MOTA	MOTP	$\mathcal{T}_P$	$\mathcal{F}_P$	$\mathcal{F}_N$	Id_{sw}
DPF-ACF	.54 / .01	.53 / .00	.87 / .01	.34 / .11	.09 / .01	76.3 / 7.8
Hierarchy-ACF	**.88** / .00	**.68** / .00	**.93** / .00	**.05** / .00	**.05** / .00	79.1 / 0.0
RJMCMC-ACF	.73 / .03	.65 / .00	.83 / .02	.10 / .02	.16 / .02	**59.0** / 19.6
DPF-DPM	.68 / .01	.53 / .01	.77 / .01	.09 / .01	.21 / .01	55.9 / 12.0
Hierarchy-DPM	**.86** / .00	**.70** / .00	**.90** / .00	.04 / .00	**.09** / .00	55.8 / 0.0
RJMCMC-DPM	.61 / .01	.59 / .00	.62 / .01	**.01** / .01	.37 / .01	**49.8** / 8.8
DPF-HOG	.52 / .01	.30 / .00	.81 / .04	**.25** / .01	.17 / .03	**12.1** / 2.6
Hierarchy-HOG	**.88** / .00	**.65** / .00	**.92** / .00	.04 / .00	**.08** / .00	37.2 / 0.0
RJMCMC-HOG	.54 / .03	.46 / .01	.61 / .01	.07 / .02	.31 / .01	56.8 / 6.1

Table 4: CLEAR-MOT results on the PETSS2L1 dataset. All results reported as μ/σ computed over five runs.

Table 4 details the results obtained on the PETS2L1 dataset. In this dataset, all detectors provided high detection and precision rates. Accordingly, an 88% MOTA is acquired with the Hierarchy-HOG tracker. The maximum intra-detector MOTA variation with the Hierarchy based trackers is approximately 2%. In this case, the Hierarchy combined with any of the detectors shows better accuracy and precision compared to the other trackers. The worst result, an accuracy of 52% is obtained with DPF-HOG.

	Hierarchy			DPF			RJMCMC		
	ACF	DPM	HOG-SVM	ACF	DPM	HOG-SVM	ACF	DPM	HOG-SVM
MOTA	.23/.69	**.67**/.25	.50/.33	.18/.32	.65/.05	.29/.20	.35/.37	.56/**.16**	.40/.21
Average		**.45**/.45			.38/**.29**			.42/**.22**	

Table 5: MOTA of tracker-detector combination averaged over the different datasets. Results are reported as μ/σ. The overall average MOTA is also indicated for each tracker variant.

Dataset	MOTA			ID Switch (average count)		
	Hierarchy	DPF	RJMCMC	Hierarchy	DPF	RJMCMC
EnterExit	.43/.28	.38/.25	**.45/.12**	**9.7/7.3**	11.5/8.3	16.2/14.8
OneShop	.06/.49	.15/.41	**.19/.20**	110.9/79.2	**12.0/5.2**	41.9/29.9
PETSS2L1	**.87/.01**	.59/.07	.63/.10	57.4/21.0	**48.1**/31.0	52.7/**9.7**

Table 6: Average MOTA and ID switch counts of the different trackers (averaged over detector combinations) on the three datasets. Results are reported as μ/σ.

For better analysis, we report two summarized results in tables 5 and 6. Table 5 reports the μ/σ of the MOTA obtained by the tracker-detector combination across the three datasets and the overall average MOTA, across dataset and detector. Clearly, the best result is obtained using the Hierarchy based trackers. When combined with DPM, Hierarchy-DPM, this leads to an average 67% MOTA across all the datasets. This result is trailed by the RJMCMC, and then the DPF. Similarly, table 6 details the average MOTA and Id switch average across the different detectors on each dataset. On the EnterExit and OneShop dataset, the RJMCMC on average does better than the rest, whereas on the

PETSS2L1, the Hierarchy based tracker does better. In terms of Id switch, the DPF does better on the OneShop and PETSS2L1 datasets, whereas the Hierarchy on average does better on the EnterExit dataset.

4. Discussions

As we can see in Tables 2, 3 and 4, the Hierarchy tracker has the best results in two out of three datasets. Its performance varies with the associated detector. Indeed, when a detector has a high precision/recall ratio (in a sequence) compared to the others, the Hierarchy tracker will get a better accuracy when combined with it than with the others. Therefore, even with a complex target representation model, the Hierarchy tracker is only showing improvements when combined with a reliable detector. Meanwhile, if the performance of two detectors are comparable, the accuracy of this tracker will not be affected much. The opposite is also true as the results of the tracker changes significantly with high variation in the detector performance. Moreover for the Hierarchy tracker, it can be inferred from tables 5 and 1, the average MOTA across datasets exhibits a 45% standard deviation with only, on average, a 6.8% and 11.1% standard deviation in detector recall and precision respectively.

The RJMCMC tracker is overall the second best filter behind the Hierarchy as it has the best results on one over three datasets. Even with its simple target representation (compared to the others) it is more resilient to the detector performance variation – better filtering capabilities. This is due to its target interaction model which seems more relevant and leads to better results (seven out of nine tests compared with the particle filter). In fact, as we can see in Table 5, it has 11% better average MOTA than the DPF. Besides, when the detector performance deteriorates, the RJMCMC has an accuracy higher than the Hierarchy filter (three out of five cases). The DPM has the best accuracy in two out of nine test cases which rates it behind the two other filters. These results are justified by the fact that the particle filter has a simple target representation and no target interaction model. Moreover, it has a better accuracy variation than the Hierarchy filter across the detector performance change and across the dataset change.

MOTP is also affected by the utilized detector. Indeed, a higher precision of the detector in a given sequence will increase the ones of the trackers'. We can notice that the DPM has the best precision, followed by the ACF and finally the HOG-SVM. Actually, the HOG-SVM has a fixed bounding box scale ratio which lowers its precision as we can see in Tables 2, 3 and 4. In addition, Table 6 show that the DPF is the best in terms of Id switches followed by the RJMCMC and the Hierarchy tracker. If the particle filter is the best, it is because it creates less tracks than the other two filters which helps it have less Id switches. Moreover, even though the Hierarchy tracker has a more complex target representation, the target interaction model puts the RJMCMC in front the Hierarchy tracker in terms of Id switches.

5. Conclusions and future works

In this work, we have evaluated three 'tracking-by-detection' approaches with three different detector combinations on three public datasets. The results show that the overall performance depends on how challenging the dataset is, the performance of the detector on the specific dataset, and the tracker-detector combination. The choice of the exact detector to use in 'tracking-by-detection' should be carefully investigated, and if possible verified on a validation set before plugging into a tracker as state-of-the-art detector does not necessary lead to better detection performance (ACF vs DPM). A tracker using a rich target representation model, as in the Hierarchy, a precise detector, as in DPM, and utilizing an interaction model, as in the RJMCMC, will perform significantly better across datasets of varying challenges.

As a future work, we aspire to broaden the evaluation to incorporate more 'tracking-by-detection' approaches and more public datasets. In addition, further investigations would be oriented in coupling the RJMCMC with rich target representation model inspired from the many trackers highlighted in [11].

References

[1] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1:1–1:10, January 2008.

[2] M. Breitenstein, R. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE T-PAMI*, 33(9):1820–1833, Sept 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR)*, San Diego, CA, USA, June 2005.

[4] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE T-PAMI*, 36(8):1532–1545, 2014.

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE T-PAMI*, 34(4):743–761, 2012.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010.

[7] D. G. Gomez, F. Lerasle, and A. M. L. Peña. State-driven particle filter for multi-person tracking. In *ACIVS*, pages 467–478, 2012.

[8] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE T-PAMI*, 27(11):1805–1918, 2005.

[9] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans. In *IEEE CVPR*, pages 1–8, June 2007.

[10] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, Mar 2004.

[11] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *IEEE CVPR*, 2013.

[12] J. Zhang, L. L. Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *IEEE AVSS*, pages 379–385, Sept 2012.