Collaborative tracking for multiple objects in the presence of inter-occlusions

Xiao, Jingjing; Oussalah, Mourad

DOI:
10.1109/TCSVT.2015.2406193

Collaborative Tracking for Multiple Objects in the Presence of Inter-occlusions

Jingjing Xiao, *Student Member*, Mourad Oussalah[1]

***Abstract*— In this paper, a new collaborative tracking algorithm is put forward to track multiple objects in video streams. First, we suggest a robust color-based tracker whose model is updated by online learned contextual information. A recursive method is performed to improve the estimation accuracy and the robustness to cluttered environment. Then, we extend this tracker to multiple targets. In order to avoid the problem of ID switch in long term occlusion, we design a hierarchical tracking system with different tracking priorities. First, the algorithm employs an adaptive collision prevention model to separate the nearby trajectories. When the inter-occlusion happens, the holistic model of tracker splits into several parts, and we use the visible parts to perform tracking as well as occlusion reasoning. In case where the targets have close appearance models, a trajectory monitoring approach is employed to handle the occlusion. Once the tracker is fully occluded, the algorithm will re-initialize particles around the occluder to capture the re-appeared target. Experimental results using open dataset demonstrate the feasibility of our proposal. Besides, comparison with several state of arts trackers has also been performed.**

***Index Terms*—particle filter, multiple targets, collaborative tracking.**

I. INTRODUCTION

AUTOMATIC tracking becomes increasingly important as thousands of low-cost and small-scale image sensors have been used to deploy surveillance systems across large cities, which renders any manual check of videos a very expensive task [1]. This motivates the extensive work carried out in this area in the last two decades as suggested by recent overview paper [2]. Usually, these tracking algorithms mainly include three primary components: target representation, spatiotemporal prediction and model update [3].

Many different cues can be used for representing the targets [2], i.e., color histogram [4], histogram of oriented gradients (HoG) [5], covariance region descriptor [6], and Haar-like features [7]. These features can be utilized in two distinct schemes: holistic approach [8-9] where a single appearance model for the whole target is employed, and a subspace-based approach [10-11], where different parts of the target are associated distinct appearance models.

Search mechanisms to locate the target can be classified into gradient descent-based methods like mean-shift [12] and spatiotemporal prediction where some pre-specified motion model was employed as in Kalman Filter [13]. The key in the estimation process in both cases is to match the appearance model of the region predicted by the spatiotemporal model with that of the target. Both gradient descent and spatiotemporal prediction methods were successfully implemented in many applications [3]. Nevertheless such methods are also vulnerable to local minima that can cause divergence. To alleviate this problem, stochastic search methods such as particle filters [14] have been widely employed due to their proven efficiency and computational cost. Besides, the appearance model of the target might also change over time due to changes in lighting conditions for instance, which requires a regular update. However, improper updating might result in some drifting problems. Despite a lot of effort and increased computational resources, effective and robust solution to the problem is still far from satisfactory in real scenarios [15]. The causes of tracking failures include inappropriate handling of background clutter, occlusion, illumination changes and target deformation, among others [16]. For this purpose, several approaches have been proposed. Some of these proposals are based on the concept of mining local visual information surrounding the target. For instance, Wang et al. [27] on-line updated the appearance model by selecting the discriminative features with the aid of existing background particles. Work in [18] advocates the use of contextual information like approach, especially in case of occluded targets. This yields robust individual trackers.

In order to deal with the problem of tracking multiple objects, the simplest way is likely to use multiple independent trackers (M.I.T.) where each target is associated to an independent tracker regardless of the state of other targets. Such trackers, e.g. [19], do not need a pre-training (beyond an initial bounding box) and often yield locally optimal solutions that best-match the target models in clear scenarios of absence of strong overlapping, lighting conditions and shape deformations. Such approach has been employed with some success to follow multiple hockey players [20] and track multiple people on the ground [21] in real time. Alternative approach is to use a pre-trained detector to scan all frames of a video, and then link successive detection instances by "tracklets" [22, 23], which is often formulated as a linear assignment problem where the cost of linking one tracklet to another is expressed as a function of appearance and motion features. Nevertheless as soon as an occlusion occurs or targets have (almost) the same appearance, M.I.T approach is acknowledged for its limitations and may result in an identity switch [24]. Qu.et al. [37] explicitly considered the interaction between multiple

[1] The Authors are with University of Birmingham, School of Electronics, Electrical and Computer Engineering, Edgbaston, Birmingham, B15 2TT, UK. Emails: jxx278@bham.ac.uk; M.Oussalah@bham.ac.uk

targets with a designed magnetic-inertia potential model. However, such model may degrade the overall performance when the targets already have distinguished appearance. Moreover, the proposed method can hardly handle the long-term full occlusion problem, i.e. the target moves together with occluder.

In this paper, to cope with the problems mentioned above, our proposed algorithm first builds a robust individual tracker that can accommodate a clutter and appearance change environments. For this purpose, first a general framework of color based particle filtering tracking is adopted [14]. Second, one utilizes the contextual information to allocate the confidence of the target's appearance changes. Third, in the resampling stage, the particles are selected according to the weights computed from their appearance similarity scores. Next, inspired by work of [16], this robust individual tracker is extended to multiple trackers through integrating a collision prevention model, which prevents tracker *jump-over* scenario. Besides the appearance similarity scores are employed to adaptively control the strength of this collision prevention power. If the targets share (almost) the same appearance model, a trajectory based monitoring strategy is employed in order to discriminate the trackers. Unlike [16], which rather employs a holistic like approach, a subspace-based method is advocated in this paper as soon as a possible overlapping scenario is detected. This allows us to handle more efficiently collision scenarios in which only part of object (target) is visible so that a contextual occlusion reasoning followed by a hierarchical tracking priority based approach are employed to distinguish the trackers. If a full occlusion occurs, an approach similar to that in [15] is adopted. Namely, based on the concept of "object permanency" which suggests that a fully occluded target will re-emerge from its occluder, the algorithm will re-initialize the particles around the occluder so that the tracker can capture the target, once reappeared, immediately.

In overall, this paper expands the original color based particle filter with original extensions that can be summarized in the following:

- In order to deal with possible occurrence of outlier particles that can shift the overall estimation, a recursive estimation, which reweights and reinitializes outlier particles, is introduced.
- The contextual information is employed to update the reference model, which overcomes the luminosity change and background influence.
- In order to tackle possible occlusion in case of multiple object tracking, a collision prevention model is introduced by reweighting particles when the target becomes close enough to each other.
- A mechanism for identifying and tackling partial and full occlusion are put forward.
- An extensive comparison of the proposal with state of art trackers using some publicly available dataset is carried out in order to demonstrate the feasibility and effectiveness of the proposal.

The rest of the paper is organized as follows. Section II details our individual tracker. The cooperation tracking scheme is described in Section III. Section IV gives details about the experiment while highlighting some of the results. A further analysis and conclusion are drawn in the Section V.

II. SINGLE ADAPTIVE COLOR-BASED TRACKER

A. Background

First, one shall mention that in our model, the target is modelled by a rectangular region, corresponding to the bounding box of the target; namely,

$$L = \{P^x, P^y, H^x, H^y\} \quad (1)$$

Where P^x, P^y represent the x-y coordinates of the center position of the bounding box, H^x, H^y stand for the region width and height, respectively as described in Fig.1.

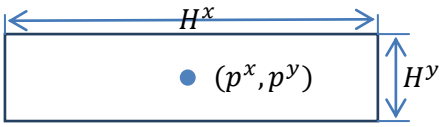


Fig. 1. Tracker's bounding box

At frame k, the tracker is represented by the state vector $X_k = \{L_k, A_k\}$ where L_k describes the attributes of the target as in (1) and A_k is a vector describing the appearance model (RGB color histogram) in the region specified by L_k. The basic tracking scheme shares the same spirit with the state-of-the-art Particle Filter tracker [14] which utilizes a set of weighted particles to represent the posterior density function associated to state vector variable L_k. More specifically, for the i[th] particle characterized by a state variable L_k^i and a weight $\omega_k^{(i)}$, a (fixed) motion model is used to propagate the state variable as:

$$L_k^i = ML_{k-1}^i + v_{k-1} \quad (2)$$

Where M stands for a constant motion matrix (unit matrix), and $v_k \sim \mathcal{N}(0, R)$ is a zero-mean Gaussian noise with a constant variance-covariance matrix R.

The weight $\omega_k^{(i)}$ of the i[th] particle is measured by the similarity between the observed appearance model and the reference target:

$$\omega_k^{(i)} = e^{\lambda_a B(A_k^{(i)}, A_k^{ref})} \quad (3)$$

Where λ_a is a constant parameter, $A_k^{(i)}$is the observed appearance of particle i while A_k^{ref} is the appearance of the reference target. $B(.,.)$ measures the similarity score through Bhattacharyya distance [25]. These weights are then normalized (by division by the sum $\sum_{i=1}^{N} \omega_k^{(i)}$).

The parameter λ_a controls the order of magnitude of the weights $\omega_k^{(i)}$. A high value of $\lambda_a$would ultimately yield high values of weights even for distances which are relatively high, increasing the danger of particle degeneracy. While a too small value of the parameter yields a compact set of particles. Consequently, λ_a should compromise between the risk of losing the target because of either particle degeneracy or compactness. In our study, setting value λ_a=5 seems to work well.

The estimation of the target from all particles is computed by averaging over the set of all particles:

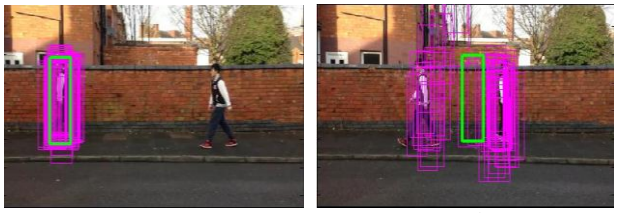
$$\hat{L}_k = \sum_{i=1}^{N} \omega_k^{(i)} L_k^{(i)} \quad (4)$$

The update of the reference model is implemented by the equation:

$$A_{k+1}^{ref} = (1 - \lambda_r) A_k^{ref} + \lambda_r A_{k+1} \quad (5)$$

Where λ_r is some constant parameter that tradeoffs the current appearance with previous reference appearance estimate. In other words, the updated reference appearance model is constructed as a convex combination of the previous estimate of the reference appearance and the current observation of the target appearance. In the absence of prior knowledge about the change of reference appearance model, setting λ_r=0.5 seems appropriate.

However, estimation (2-4) might not be fully accurate, especially in a cluttered environment. Indeed, for instance, when the target encounters a similar background, the estimation induced by some particles might be fully erroneous, so that the distribution of particles yields distinct clusters, which, in turn, leads to an overall target estimate according to (4) located somewhere between the clusters but far away from the target, as it can be seen in Fig. 2. Similarly, a slight inaccuracy in updating stage (5) can be accumulated over time, and may yield a serious drifting, which causes target divergence. In Fig. 2, the purpose is to track one specific individual (target object) in a video containing two individuals who look almost alike, yielding a close appearance model too, and walking in opposite directions. Initially, before the crossing, the estimation looks pretty consistent as demonstrated by a consistent set of bounding boxes related to various particles. However after crossing, on the right hand side of Fig. 2, the estimation as depicted by the thick-lined square is very much biased by the second person in the image.



a) Frame 1 b) Frame 55. After cross-over

Fig. 2. Failure modes in environmental clutter

This motivates our proposal of a recursive estimation method to solve the problems in clutter environment where contextual information is extracted to form a robust model updating. This is detailed in the next section.

B. Recursive estimation

As pointed out in previous section, the main cause of drift observed in the presence of clutter relies on the fact that particle estimates are mainly distributed across several clusters, while some particles act as outliers, which yields an erroneous global target estimate. The key idea pursued in this paper to handle this issue is to introduce a new weighting of the particles, which takes into account the actual (global) target estimate and iterates until no outliers is detected. Intuitively this assumes a smooth transition of target estimate from one frame to the next one. More specifically, let $d_k^{(i)}$ be the distance from the i[th] particle bounding box center of the k[th] frame to the target estimate (computed as average over all associated particles) at previous frame:

$$d_k^{(i)} = \sqrt{(P_{k-1}^x - (P_k^x)^i)^2 + \left(P_{k-1}^y - \left(P_k^y\right)^i\right)^2} \quad (6)$$

The average distance across all particles is therefore:

$$\bar{d}_k = \frac{1}{N} \sum_{i=1}^{N} d_k^{(i)} \quad (7)$$

A particle i is considered drifting away from the normative estimation if and only if:

$$d_k^{(i)} > \lambda_d \bar{d}_k \quad (8)$$

λ_d is a constant parameter related to the variability of frames in the underlying video. Especially, high value of λ_d reduces the number of outliers while setting its value less than one would substantially increase the number of outliers. Therefore a compromise situation, which avoids high values that would prevent detection and very small values that would yield false positives, is required. Setting $\lambda_d = 2$ is found to work well in our case.

The particles will then be resampled in the light of constraint (8), which would avoid large discrepancy of particle distribution. Figure 3 illustrates the above resampling scheme. The plot displays a configuration in which two particles, say, p[th] and q[th] particle, act as outliers with respect to the rest of the particles, which rather forms an homogeneous group, yielding a global target estimate $\hat{L}$ shifted away from the homogeneous group of particles ($\hat{L}\_$ stands for estimate at previous frame). The two outlier particles are assumed to fulfil the constraint (8).

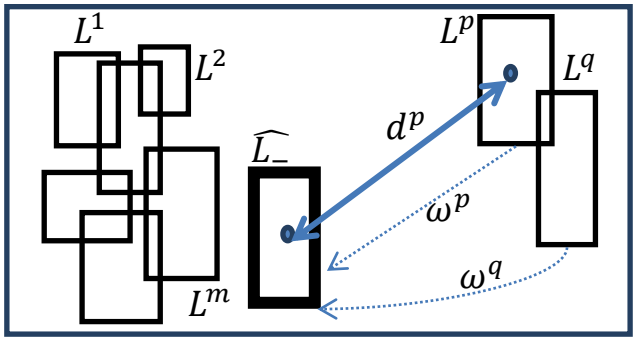


Fig.3. Resampling scheme

The new resampling strategy works as follows:

(i) Initialize each outlier target estimate to the global estimate $\hat{L}$,

(ii) Calculate the appearance model in the region delimited by $\hat{L}$

(iii) Compute the new weights attached to the p[th] and q[th] particles according to (3); namely, using the notations of Fig. 3,

$$L'_p = \hat{L} \text{ and } L'_q = \hat{L} \quad (9)$$

$$\omega^{(p)} = \omega^{(q)} = e^{\lambda_a B(A^{(\hat{L})},\ A^{ref})} \quad (10)$$

where $A^{(\hat{L})}$ corresponds to the appearance associated to the global estimate $\hat{L}$;

(iv) Renormalize the weights accordingly, yielding $\omega'^{(i)} = \omega^{(i)} / \sum_k \omega^{(k)}$.

(v) Compute a new global estimation according to (4).

(vi) Repeat steps (6-8) to find out whether there are any outlier particles.

The resampling steps (i)-(v) are iterated until no outlier is generated. Besides, in practice the number of the above iterations is substantially reduced as the number of outliers tends to stabilize after the first resampling stage. A pseudo-code summarizing such resampling is described in TAB.I.

TABLE I. PSEUDO-CODE OF RECURSIVE ESTIMATION

Recursive Estimation- Resampling. Input: $\{L_k^{(i)}, \omega_k^{(i)}\}$

a. Estimate the global state $\hat{L}_k$ using Eq.4.

b. **FOR** Each particle i

Compute the distance $d_k^{(i)}$ using Eq. (6)

END

c. Compute average distance $\overline{d_k}$ using Eq. (7)

d. While particle j fulfils Eq. (8), **DO**

- Set $L_k^j = \hat{L}_k$
- Compute appearance model $A_k^{(\hat{L})}$
- Compute new weight as $\omega^{(j)} = e^{\lambda_a B(A_k^{(\hat{L})}, A_k^{ref})}$

END

e. **FOR** Each particle i

Normalize weights using $\omega^{(i)} = \omega^{(i)} / \sum_k \omega^{(k)}$.

END

f. **GO** to a)

g. Repeat a)-f) until no particle j fulfils (8)

h. Output global estimation $\hat{L}_k$, and particles $\{L_k^{(i)}, \omega_k^{(i)}\}$

C. *Contextual information updating*

Updating the reference model is often problematic because of difficult predictability of future lighting conditions or so in video sequences. The convex combination between previous model and current appearance of target in view of expression (5) might be misleading because of the possible influence of the background, which is ignored in the target appearance model. Inspired by work of Talha and Stolkin [26] and Wang et al. [27], we first suggest to account for the background through enlarging the estimated target's bounding box by a fixed proportion. Second, the multiplicative factor λ_r in (5) is chosen to account for the proportion of background pixels with respect to foreground pixels at each bin interval of the (intensity) histogram. The rational for doing so is that by comparing the appearance change between the model estimation (foreground) and the background of the (learned) reference, if the changes are quite similar to the background, a low factor will be assigned, indicating the prevalence of the previous reference model estimate in (5). Otherwise, the contribution of the current foreground should be made more important. More formally, the bounding box associated to target estimate of the reference is enlarged uniformly as in Fig. 4 so that a local background region defined as the complement of the foreground region with respect to the aforementioned enlarged region is estimated.

Specifically, we enlarge the bounding box of target estimate by a constant scaling factor $\sigma > 1$ for each edge as shown in Fig. 4 so that the size or resolution of the enlarged area becomes:

$$S_k^{f+b} = \sigma^2 H_k^{*x} H_k^{*y} = \sigma^2 S_k^f \tag{11}$$

The appearance model (histogram) A_k^{f+b} of the enlarged bounding box S_k^{f+b}, which contains both foreground and background information, can be derived straightforwardly. Next, the histogram of the background region only A_k^b is computed using:

$$A_k^b = \frac{A_k^{f+b} S_k^{f+b} - S_k^f A_k^f}{S_k^{f+b} - S_k^f} \tag{12}$$

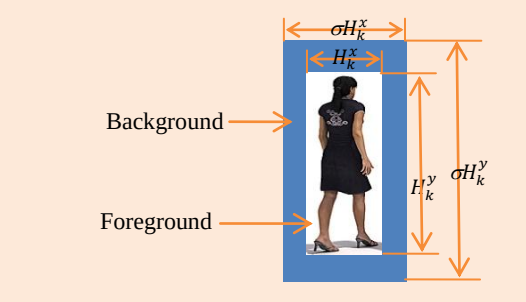


Fig.4 Foreground and Background region

Note that at each histogram bin u, the quantity $A_k^{f+b}(u) S_k^{f+b}$ represents the total number of pixels of the S_k^{f+b} region whose grey values fall within the interval delimited by bin u. Consequently, the quantity $A_k^{f+b}(u) S_k^{f+b} - S_k^f A_k^f(u)$ corresponds to the total number of pixels in the background region S_k^b whose grey level values fall in the interval delimited by bin u. The denominator in expression (12) allows us to normalize A_k^b within unit interval.

Following [26-27], we set the multiplicative factor σ=1.2. Notice that this choice can also be influenced by the intensity of the clutter and level of interactions among targets, if any, as we may end up with a background which also contain possible other targets. Consequently cautious is needed when sequence of multiple targets is used by reducing σ. Nevertheless it should always be held $\sigma > 1$.

Next, the counterpart of λ_r in Eq.5 is determined for each bin as:

$$C_k(u) = \begin{cases} 1 - e^{-\lambda_c \left(\frac{A_k^f(u)}{A_k^b(u)} \right)}, & A_k^b(u) \neq 0 \\ 1, & A_k^b(u) = 0 \end{cases} \tag{13}$$

Where λ_c is a control parameter that quantifies any preference of foreground model over background model. Motivated by a cautious attitude when background model is attributed higher weight as in alternative studies, we set λ_c =0.1, which is proven to work well in practice. Typically, it is easy to see from (13), that the more dominant the foreground model with respect to background model, the higher the weighting factor $C_k(u)$, which, in turn, makes the contribution of the previous reference model in (5) more important. In other words, if the pixels, whose grey level intensities belong to the given bin u, are dominantly located in the foreground or the estimated bounding window, the influence of the background is negligible, and, therefore, there is no need to influence much the reference model, yielding a

reference model close to its predecessor (in previous frame). Otherwise, the background is deemed to be important and, therefore, the reference model should be changed accordingly, yielding a smaller coefficient factor for A_k^{ref}, but higher for current appearance model A_{k+1}. Also, notice that distinguishing the case $A_k^b(u) = 0$ in (13) ensures the continuity of C_k with respect to u, while its values range in the unit interval. A counterpart of (5) can be written as:

$$\hat{A}_k^{ref\prime}(u) = C_k(u)A_{k-1}^{ref}(u) + \big(1 - C_k(u)\big)A_k^f(u) \quad (14)$$

The newly calculated appearance model (14) should also be normalized if any (by dividing by the sum over all bins). A generic pseudo code of individual tracker is summarized in TAB.II.

TABLE II. INDIVIDUAL TRACKER

Individual tracker: Adaptive color-based tracker
Given the sample set $\{L_0^{(i)}\}$ and the target model. Perform the following steps: 1. **Predict** each sample from the set $\{L_{k-1}^{(i)}\}$ by a linear stochastic differential equation. Eq.2. 2. **Observe** the colour distribution in the region $\{L_k^{(i)}\}$, and calculate the weights of the particles $\{\omega_k^{(i)}\}$. Eq.3 3. **Recursive estimation**. Output global estimation $\hat{L}_k$, and particles $\{L_k^{(i)}, \omega_k^{(i)}\}$ as in TAB. I. 4. **Update** the reference model according to the contextual information. Eqs.13-14.

III. COLLABORATIVE TRACKING

A. *Introduction*

Extension of individual tracker to track multiple objects is not straightforward as frequent interactions between such objects not only bring heavy occlusion problem but also the risk of *identity-switch*. As pointed out in the introduction of this paper, the use of independent individual trackers as a way to deal with multiple object tracking is not very effective. Instead, mechanisms for monitoring all pairs of (particle) target estimation in order to avoid possible occurrences of occlusion are required. This enables what we refer here by *collaborative tracking*. In the latter, the distance between any pairwise target estimates is constantly monitored. In this respect, four distinguished cases can be reported:

- If the distance is sufficiently large, then the rationale is to use multiple independent trackers (M.I.T), indicating the absence of any occlusion or target identity switch.
- If such distance is smaller than some predefined threshold but without causing an overlap of the two bounding box estimates, then an adaptive prevention model will be applied, where the distance is explicitly taken into account in refining the likelihood (weight).
- If there exists an overlapping between the two target estimations, then a (partial) occlusion-based reasoning will be enabled in order to distinguish the occluded target from the non-occluded one, and then refine the estimation accordingly.
- If the distance indicates a full occlusion, e.g., one bounding box region is fully included into the other one, then the full occlusion reasoning is activated, where basically, one waits for the re-appearance of the target.

The above constitutes our collaborative tracking for dealing with multiple targets, where the various subcases are detailed in the subsequent subsections.

B. *Adaptive collision prevention model*

The main idea in collision preventive model is that as soon as the distance between the (global) estimates of the two targets is less than some predefined threshold and the regions are non-overlapping, then the weights of the particles in the next frame will be refined to take into account both the distance to the other target as well as the dissimilarity of the appearance models of the two targets.

More formally, given two targets X_1 and X_2 (objects in the video that one wants to track), with particles $\{L_{X_1(k)}^{(i)}, \omega_{X_1(k)}^{(i)}\}$ and $\{L_{X_2(k)}^{(i)}, \omega_{X_2(k)}^{(i)}\}$ i=1,N, respectively. The current weight $\omega_{X_1}^{(j)}$ of the j[th] particle of target X_1 will be refined as (omitting the subscript k for simplicity of notations)

$$\widehat{\omega}_{X_1}^{(j)} = \omega_{X_1}^{(j)} exp(\lambda_{X_1X_2} \cdot d(R_j^{X_1}, R^{X_2})) \quad (15)$$

Where $R_j^{X_1}$ and R^{X_2} are bounding box regions associated to j[th] particle of target X_1 and global estimate of target X_2, respectively. $d(.,.)$ is the distance between the two bounding boxes in the sense of minimum distance between corners of the two regions.

Similar reasoning applies to particle j of target X_2 where the counterpart of (15) is:

$$\widehat{\omega}_{X_2}^{(j)} = \omega_{X_2}^{(j)} exp(\lambda_{X_1X_2} \cdot d(R_j^{X_2}, R^{X_1})) \quad (16)$$

The coefficient factor $\lambda_{X_1X_2}$ determines the similarity of the appearance models associated to the two targets, and is given by

$$\lambda_{X_1X_2} = \lambda_d B(A^{X_1}, A^{X_2}) \quad (17)$$

Where λ_d and $B(.,.)$ are defined as in expression (3). Weights in (15-16) are also normalized in similar way. Trivially from (15-17), the weight of particle is increasing with respect to both the distance between the two target estimates and the dissimilarity of their appearance models at the target estimates. This translates the fact that particles should contribute less to the global estimate of the target when they are sufficiently close to the other target. On the other hand, if the two targets in terms of their global estimates have the same appearance models, the quantity $B(A^{X_1}, A^{X_2})$ coincides with unity, making the $\lambda_{X_1X_2}$ constant. Otherwise, the more distinct the appearance models of the two targets, the higher the associated weight of the particle. This is also in agreement with the intuition that distinguished particles in terms of appearance models should contribute more to the global estimate, since close appearance models increase the chance of identity-switch phenomenon.

C. *Hierarchical tracking priority (in case of partial occlusion)*

Throughout this section one considers situation of a partial overlapping of the bounding box regions associated to the two targets X_1 and X_2. In such case, the main research questions that are handled here are:

- Does the overlapping of the bounding box regions entail a partial or a full occlusion of targets?
- Which target is occluded in another one?

- How to refine the estimation of the targets in case where occlusion is confirmed?

Strictly speaking, the occurrence of such overlapping does not necessarily entail the occurrence of occlusion. For instance, in the case of a minor overlapping, it cannot be excluded that the actual shapes of the two targets are completely identified without any ambiguity. On the hand, it also holds that the overlapping may rather be due to a wrong estimation instead of a genuine occlusion, so that one of the bounding box regions may not include any genuine target. In such cases, the occurrence of overlapping does not imply any occlusion. Similarly, even if the occlusion is confirmed, still one requires to determine whether target X_1 is occluded by target X_2 or vice versa. At each case, an appropriate reasoning for estimating the target will be enabled.

For this purpose, the use of appearance model of the two targets sounds rational. However, if the two targets have similar appearance models, then alternative reasoning will be required where monitoring each target itinerary will be employed as detailed later on. Especially, the itinerary-based monitoring will be triggered as soon as it holds that

$$B(A^{X_1}, A^{X_2}) \geq T_a \quad (18)$$

Expression (18) indicates that the appearance models between the two targets is deemed to be similar as soon as their similarity values in the sense of Bhattacharya coefficient is larger than some predefined threshold T_a. Here T_a=0.95 was used. Ideally, T_a close to one indicates a strict equal similarity in the sense of Bhattacharyaa distance, a situation which may rarely occur in practice, while relaxing the value of T_a allows us to account for targets of close appearance models.

a) Targets with different appearance models

i) Occlusion confirmation

Once there is an overlap between the target estimates (bounding box regions), and a clear distinct appearance models, we propose a two-step competition mechanism in order to confirm or refute the existence of occlusion. First, the likelihoods ω_{X_1} and ω_{X_2} of both targets in the sense of expression (3) are computed using the appearance model A^{X_1} and A^{X_2} around the (global) estimated bounding box of each target and using the information on the reference model A^{ref} (for both targets). The conjecture is that if there is an occlusion then it holds that one of the likelihoods will be dominant. Nevertheless, the discrepancy between the values attached to the two likelihoods can also be due to a wrong target estimation. Second, in order to differentiate between the cases where the discrepancy of the two likelihoods is due to a wrong estimation and that due to an occlusion, the contextual information will be employed as in Section II.C. More formally, let us assume without loss of generality that

$$\omega_{X_1} < \omega_{X_2} \quad (19)$$

(Likelihood associated to target X_1 is weak compared to that of X_2). We first calculate the positive contribution of the appearance model of another target X_2 (potential occluder) over the X_1 reference model, which can be quantified by:

$$h_1^{X_1} = \sum_{i,[A^{X_2}(i) - A^{ref_{X_1}}(i)] \geq 0} \left(A^{X_2}(i) - A^{ref_{X_1}}(i) \right) \quad (20)$$

Next, we repeat the same reasoning with the enlarged region of target X_1, and focusing on background region, yielding:

$$h_2^{X_1} = \sum_{i,[A^{X_1^b}(i) - A^{ref_{X_1}}(i)] \geq 0} \left(A^{X_1^b}(i) - A^{ref_{X_1}}(i) \right) \quad (21)$$

Where $A^{X_1^b}$ stands for the histogram associated to background region of the enlarged bounding box pertaining to target X_1. Comparing (20) and (21), if $h_1^{X_1}$ is dominant over $h_2^{X_1}$, say,

$$h_1^{X_1} > h_2^{X_1} \quad (22)$$

then X_1 is assumed to be occluded by X_2. Otherwise, the occlusion cannot be confirmed in the current frame, as the target with lower weight in Eq.19 might come from poor estimation. So, the same reasoning will be repeated in subsequent frames to check whether the occlusion can be confirmed. The same reasoning applies in the case of $\omega_{X_2} < \omega_{X_1}$ by substituting X_2 to X_1 in (20-21).

ii) Estimation in case of (partial) occlusion

If the occlusion is confirmed (using conditions like (19) and (22)), the reasoning is first to stop enabling the (adaptive) prevention collision model and, second, to discriminate the visible parts and non-visible parts in bounding box regions of targets, and, third, to re-computes the weights accordingly. One notices that the current case of partial occlusion excludes the case of full occlusion where one bounding box has no associated visible parts. This situation will be investigated later on. To illustrate our reasoning, let us exemplify in Fig. 5 a simple case of target overlapping, which confirms a partial occlusion.

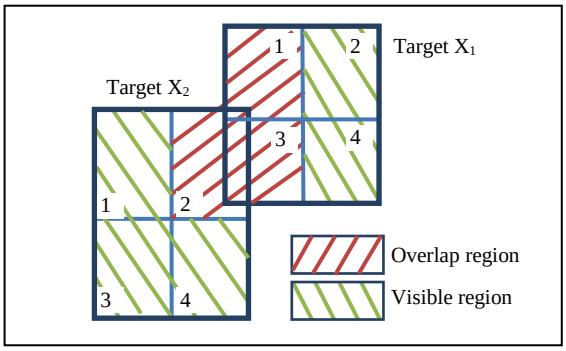


Fig.5. Overlapping of the two targets' estimates.

The approach adopted here is to create a subdivision of each bounding box region in order to account for visible / non-visible parts. Strictly speaking, the initial outcome of such occlusion reasoning is twofold:

1) A visual partition model of each target is elicited. For instance, in the case of Fig. 5, ignoring any occlusion scenario, the visible part of target X_2 corresponds to partitions 1, 3 and 4, while that of target X_1 corresponds to partition 2 and 4. This is referred to as *visual partition model*. However, if the occlusion reasoning concluded that only target X_2 is partially occluded, then target X_1

becomes fully visible, while partitions 1, 3, 4 are the visible parts for target X_2, which constitutes the associated visual partition model.

2) A new weight (likelihood) is computed for each target to account for visible parts only. More formally, for each subdivision j of the bounding box associated to target X_i, one computes the corresponding appearance model $A_{s_j}^{X_i}$, and using Eq. (3), the associated weight $\omega_{X_i}^{(s_j)}$, where symbol *s* stands for subdivision. Next, considering a set V of subdivision that belongs to visible part, the new weight of the target X_i is computed as:

$$\omega_{vX_i} = \frac{1}{|V|}\sum_{j\in V} \omega_{X_i}^{(s_j)} \quad (23)$$

With

$$\omega_{X_i}^{(s_j)} = e^{\lambda_a B(A_{s_j}^{X_i}, A^{ref}{}_{X_i})} \quad (24)$$

The subscript *v* stands for *visible* in (23). For instance, in Fig. 5, assuming only X_2 is occluded, target X_2 induces |V|=3, while there is no need to apply neither subdivision nor weight adjustment for X_1 as it is fully visible.

In the subsequent frame, the visual partition model is extrapolated to all particles of target X_1 and X_2. In other words, the estimation process is such that the weights attached to particles of both targets are adjusted according to (23-24) using the visual partition obtained in previous frame, while the tracking is performed according to M.I.T and the occlusion condition is tested again. The above is based on the assumption that the movement between two consecutive frames is small enough to justify the conjuncture of extrapolating the visual partition model.

b) Targets with same appearance

If the targets have similar appearance models, trivially, the above reasoning cannot be employed to confirm or refute the occurrence of an occlusion. The idea is therefore to monitor the trajectory of the targets and adjust the weights of the particles according to the direction of the target movement and the size of the bounding box regions. For this purpose, one requires first to determine the direction of movement of targets. Intuitively, monitoring the velocity of the center of the bounding box region within a predefined time window provides an answer to such request. More specifically, let (V_x, V_y) be the average velocity of a given tracker, computed from previous m frames (m is chosen 5 in our case), then one can use the sign of the largest absolute values between V_x and V_y to decide on the direction as it can be seen from TAB. III below.

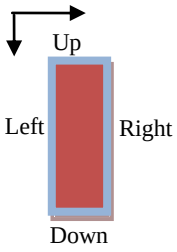


TABLE.III. DIRECTION MAP

$\max(\lvert V_x\rvert, \lvert V_y\rvert)$	Sign of V_x, V_y	Output (Direction)
V_x	+	Right
V_x	-	Left
V_y	+	Down
V_y	-	Up

Given the geometrical constraint (rectangular) of the target estimate, the direction is identified by one of the four possibilities: up, down, left and right.

Next, the idea is somehow similar to that of adaptive prevention collision model where some particles will have their weights discounted while taking into account the direction of movement as well as the position of the particles with respect to boundary case. More specifically, let us consider, without loss of generality, a situation in which target X_1 has a direction Left and overlaps with target X_2 as in Fig. 6.

In the plot l_1 and l_2 denote the left and the right edges of the tracker X_1, respectively, while δ corresponds to the length of the horizontal edge of Tracker X_2 whose bounding box region is smallest.

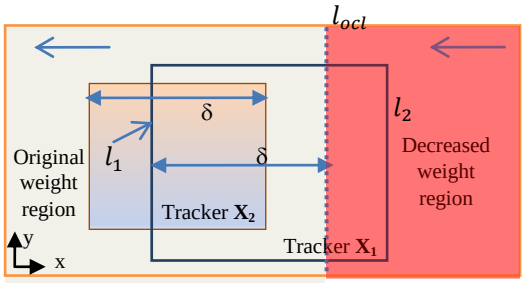


Fig.6 regions with different weights

First let us define the full occlusion using l_{ocl} (a vertical line in Fig. 6 delimiting the full occlusion scenario). Namely, as soon as l_2 coincides with l_{ocl} in Fig. 6, the bounding box of tracker X_2 is fully included in that of tracker X_1. l_{ocl} is therefore defined such that the length from l_1 to l_{ocl} is L. Second, in order to reinforce the movement of target X_1 in the left direction, all particles of X_1 located prior to l_{ocl} in the opposite direction of movement will be discounted, otherwise the weight is left unchanged. This is motivated by the conjecture that particles located far away from that separated by l_{ocl} will likely obscure the movement of the target towards the predefined direction if it was allocated higher weight. Namely, using previously employed notations and configuration of Fig. 6,

$$\omega'^{(j)}_{X_1} = \begin{cases} \gamma\omega_{X_1}^{(j)} & if\ \ P_{X_1}^{x} \le l_{ocl}, \\ \omega_{X_1}^{(j)}, & otherwise \end{cases} \quad (25)$$

Where $\omega_{X_1}^{(j)}$ is the original weight of the j^{th} particle of target X_1. γ is the discounting factor between 0 and 1, which is 0.5 in our experiment. Third, the same reasoning is repeated with target X_2, when looking at its direction and updating its particles in the subsequent frame using the counterpart of Eq. (25) for target X_2.

c) Targets re-tracking after full occlusion

Provided that the targets have different appearance models, our reasoning in case of full occlusion relies on the concept of *object permanence* which suggests that a fully occluded target will re-emerge from its occluder [15]. Besides, it is typically known that the weights associated to particles of occluding target are usually low. Therefore, in the same spirit as in [15], the idea is to randomly reinitialize the particles of the occluded target around the occluder so that the tracker can capture the reappeared target immediately after its reappearance as in Fig. 7. To confirm a re-emerge target, the tracker will compare the appearance of the newly estimated tracker to

the reference model. If the likelihood is beyond the pre-set threshold T_r, the target can confirm the re-appeared target. Setting T_r = 0.8 is found to work well in practice. Otherwise, the tracker will keep re-initializing the particles according to the position of the occluded target. The pseudo code of the algorithm is shown in the TAB. IV.

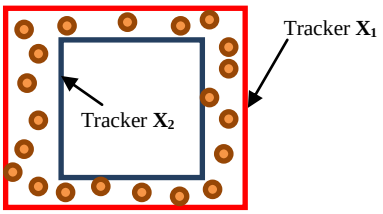


Fig.7 Full occlusion (small circles: re-uninitialized particles).

TABLE.IV. COLLABORATIVE TRACKERS

***Multiple tracker*:** Collaborative tracking in the presence of inter-occlusion
Given the state of multiple trackers $\{X_k^i\}_{i=1\ldots N}$. Perform the following steps for each pair of tracker: 1. Form the pairwise trackers $\{X_k^1, X_k^2\}$. Monitor the distance between trackers. Perform collision prevention model if distance is below threshold, otherwise, carry over using M.I.T. 2. Predict the overlap (occluded) area for both trackers, match the target with visible parts (using collision prevention model). 3. Compute the trackers' likelihood and do the occlusion reasoning; Set tracking priority to different trackers. 4. Re-initialize particles for fully occluded target.

The overall flow chart diagram of our proposed collaborative tracker is presented in Fig.8.

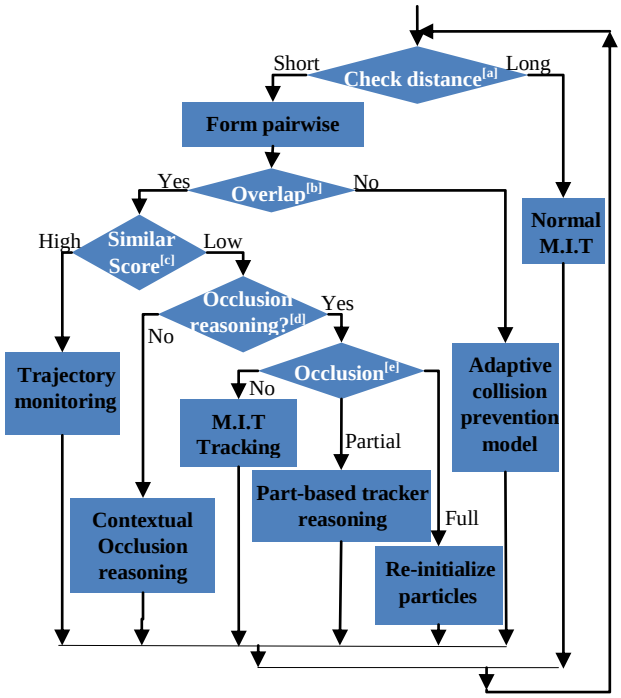


Fig.8 Flow chat diagram of collaborative tracking. [a]: Calculate the distance between the trackers associated to the two objects; [b]: check whether there is an overlap between trackers; [c]: check the similarity score of the appearance model between targets; [d]: check whether the algorithm has confirmed the occlusion relationship; [e] check whether a full occlusion is confirmed

IV. EXPERIMENTAL RESULTS

In this section, we first test our individual tracker on publicly available benchmark dataset [28]. We also employed three home-made videos, to better demonstrate the efficiency of our collaborative tracking algorithm.

A. *Evaluation metrics*

Individual tracker: In order to ease the comparison with benchmark dataset [28], the performance of individual tracker is primarily measured by two metrics: overlap and root mean square error (RMSE). The former is quantified as:

$$A_k = \frac{TT_k \cap GT_k}{TT_k \cup GT_k} \tag{26}$$

Where TT_k is the tracker's bounding box and GT_k is the ground truth bounding box. Note that if the ground truth coincides with tracker A_k=1. On the other hand, the RMSE quantifies the overall bounding box center errors between the target's predicted center TT_k^c of the tracker and the ground-truth center GT_k^c over all the frames:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} \| TT_k^c - GT_k^c \|} \tag{27}$$

Where N is the total number of frames.

Multiple trackers: By counting the number of detected objects at each frame with respect to the ground truth knowledge, the overall performance of multi-targets tracking is measured according to the following four metrics [29]:

- False Negative Ratio(FNR): $FNR = \frac{\sum_k fn_k}{\sum_k gt_k}$
- False Positive Ratio(FPR): $FPR = \frac{\sum_k fp_k}{\sum_k gt_k}$
- Miss Match Ratio (MMR): $MMR = \frac{\sum_k mm_k}{\sum_k gt_k}$
- Multiple Object Tracking Accuracy (MOTA):

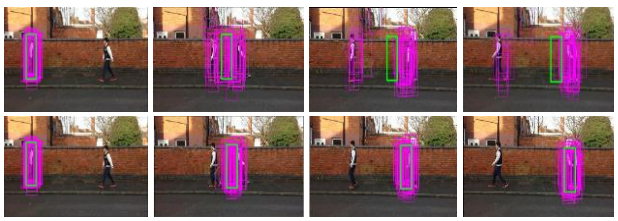
$$MOTA = 1 - \frac{\sum_k (fn_k + fp_k + mm_k)}{\sum_k gt_k} \tag{28}$$

fn_k , fp_k , mm_k and gt_k denote false negatives (misses), false positives, mismatches and ground truth at frame k, respectively.

B. *Individual tracker performance*

In the experiment, the values of some key parameters for individual tracker are provided below, which are used in all experiment:

In order to demonstrate the effect of the proposed recursive estimation, we first applied our tracker to the video shown in Fig.2. The results highlighted in Fig. 9 clearly indicate the ability of the recursive estimation (second raw in Fig. 9) to overcome the degrading influence caused by the outlier particles.



Frame 1 Frame 48 Frame 63 Frame 72

Fig. 9. Tracking experiment of recursive estimation: first raw: no recursive estimation; second raw: with recursive estimation. ***Green colour***: estimation; ***Pink colour***: particles.

Besides, in order to prove the existence of outliers in the sense of (7-8), which motivates the use of our recursive approach, Fig. 10 describes the number of outliers at each frame in case of Video shown in Fig. 2. Next, in order to demonstrate the usefulness of our tracker-update based methodology, we have chosen two challenging sequences from [28]. One corresponds to a gymnast video with important self-deformation in noisy background. The other one corresponds to a heavy illumination change.

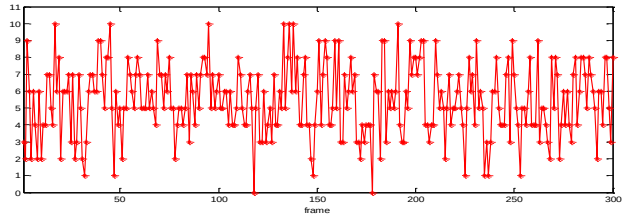


Fig. 10. The number of outlier particles detected at each frame

The study compares our approach with the standard color-based particle filter approach without update, and the one with update in [14] where the speed controlling parameter is 0.5.

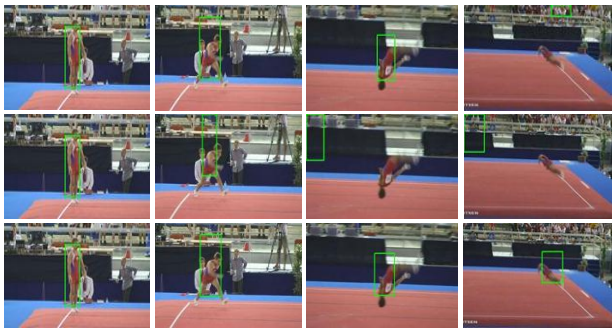


Fig.11. Tracking experiment of appearance adaption (Self-deformation): first raw: no updating; second raw: Updating in [14]; third raw: our method (frame: 1, 90, 150, 180)

In gymnast sequences (Fig.11), the athlete endures a dramatic shape deformation. The results show a loss of target at some frames in case of color-based particle filter approach without update or when the update is purely based on observation. Our collaborative tracker successfully tracks those complex cases. Strictly speaking, the absence of update in the first case induces a serious handicap to deal with abrupt variation of shape and illumination of video frames because of the lack of possibility to obtain good matching between target appearance and that of original reference, which, in turn, mostly explains the loss of targets observed in such cases. Similarly, the absence of robust mechanism to account for background clutter in [14] induces a failure. On the other hand, the use of background information in our model partially allows us to overcome such difficulty.

Fig. 12 illustrates the results of the various trackers when submitted to sharp illumination changes. In this case, again, the importance of (robust) appearance model update is noticeably stressed as tracking improvements when using such updating mechanism are clearly highlighted. While the accumulation of errors from imperfect estimation updates in [14] leads to target loss.

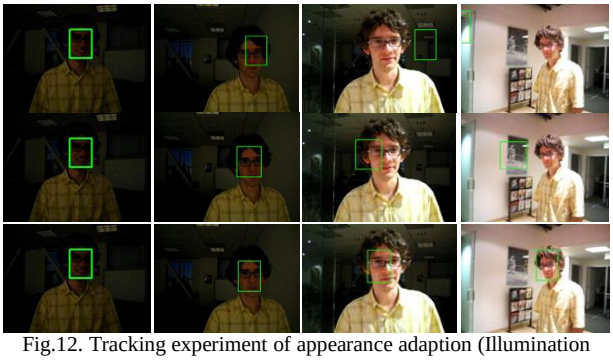
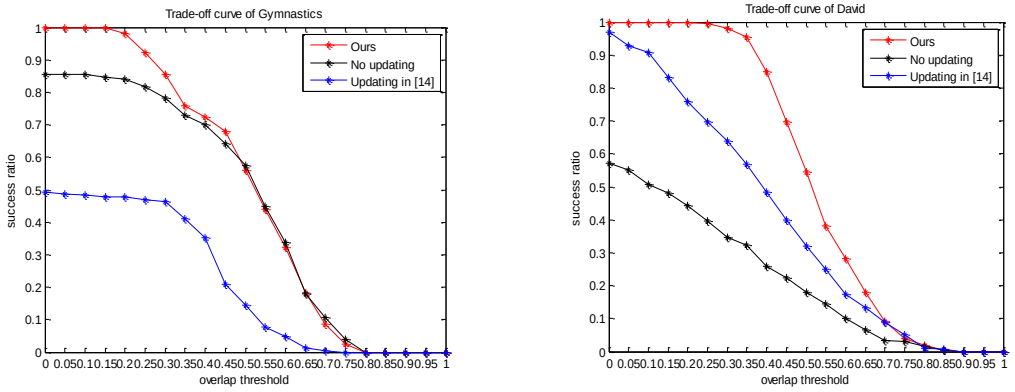


Fig.12. Tracking experiment of appearance adaption (Illumination change): first raw: no updating; second raw: updating in [14]; third raw: our method (frame: 1, 100, 200, 300)

In order to provide an overall evaluation across all frames, Fig.13 summarizes the performance of the trackers with respect to trade-off-curve. This quantifies the number of frames where the target is tracked successfully, referred as success rate, under a given overlap threshold level. Especially, with a same success rate, a higher overlap translates a better detection capability of the object by the tracker. Similarly, with the same level of overlap, a higher success rate indicates a better robustness of the algorithm. Both in case of self-deformation video and illumination change sequence, it is obvious that our algorithm clearly outperforms the other two alternative state-of-the-art methods, yielding a significantly better accuracy.



(a) Self-deformation (b) Illumination change

Fig. 13. Trade-off curve of Gym and David sequences

For a further comparison, we also tested our algorithm using some state of art sequences involving human shapes: Basketball, Crossing, Couple, David3, Subway, Walking and Woman [28]. These sequences contain challenging

occlusion scenes, low resolution, illumination change, and background clutter. Besides, five state-of-art methods, which are acknowledged for their good tracking performances in challenging scenes, have been employed for comparison purpose. The first one is the standard color based particle filter [14] which inspired our current work. The second one is the Struck method [32], which is based on Haar features and support vector machine (SVM) classification. The third one employs sparse representation and L_1 minimization approach [33] where L_1 regularized least square solution is employed. The fourth one uses circulant structure kernel (CSK) [34], which is based on application of circulant matrice theory and Fourier analysis to enhance the learning detection task. The last one is the discriminative model based tracker (VR) [17] which employs the background information. All these trackers share the common feature of use of appearance model to characterize each instance as opposite to motion or texture features. Besides, [32-34] share a sparse sampling strategy, requiring an offline training phase, although Struck method has proven to be efficient to online training as well. The results of different trackers in terms of average overlap metric and RSME value are shown in TAB. V and TAB. VI, respectively. The best and second-best trackers are highlighted using italic and underline representation, respectively.

TABLE.V. AVERAGE OVERLAP FOR EACH SEQUENCE

Name	Ours	PF [14]	Struck [32]	L1 [33]	CSK [34]	VR[17]
Basketball	***0.4797***	0.2555	0.0914	0.0320	0.0196	0.3309
Crossing	0.4256	0.3097	0.2021	0.1848	0.4790	***0.6425***
Couple	***0.6022***	0.5673	0.5362	0.4594	0.0751	0.0642
David3	***0.6084***	0.5796	0.2917	0.3770	0.4976	0.4463
Subway	0.3845	0.0873	***0.6684***	0.1597	0.1925	0.5613
Walking	0.5442	0.2956	0.4521	***0.6555***	0.5365	0.2409
Woman	0.1215	0.0716	***0.6089***	0.0539	0.1668	0.0801
Mean overlap over all sequences	***0.4523***	0.3095	0.4073	0.2746	0.2810	0.3380

TABLE.VI. AVERAGE CENTER ERRORS (RMSE) FOR EACH SEQUENCE

Name	Ours	PF[14]	Struck [32]	L1 [33]	CSK [34]	VR[17]
Basketball	***12***	107	126	148	312	68
Crossing	9	41	121	58	9	***7***
Couple	***9***	11	11	29	145	111
David1	***17***	18	106	90	56	78
Subway	***7***	145	8	150	164	16
Walking	7	79	8	***2***	7	121
Woman	106	130	***6***	192	208	136
Mean RMSEs over all sequences	***24***	76	55	96	129	77

C. *Multiple trackers performance*

a. Our own dataset

As current publicly available dataset is not perfectly good to show the significant improvement of our tracker under specific change of appearance models and occlusion, we tested our proposed collaborative algorithm on recorded videos in order to enable coherent comparative experiment [2]. Different challenges of the videos are shown in TAB.VII.

TABLE.VII ATTRIBUTES OF HOME BUILT VIDEOS

Name	Main challenges
Video 1	Two targets with *both similar size and appearance*, crossing trajectories.
Video 2	Two targets with *similar size but different appearance*, and frequent inter-occlusions.
Video 3	Two targets with *both different size and appearance*, long time inter-occlusion.

The values of these parameters are suggested to achieve relatively good performance.

Some of the visual results are shown in Fig.14. Bounding boxes pertaining to distinct targets are labelled by different line widths. In Video 1 (first raw), since the two targets have quite similar appearance, the occlusion reasoning is not trustable. Therefore, our collaborative tracker makes use of trajectory monitoring to track the two targets. Besides, our recursive method also significantly contributes to overcome the effect of outlier particles that arise in such scenarios.

Next, we tested our algorithm with object of distinct appearance models of Video 2 (second raw in Fig. 14).

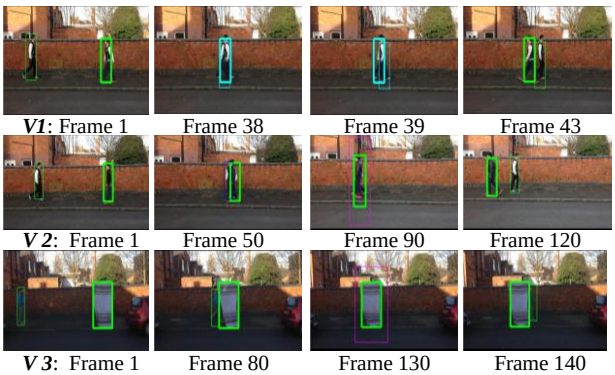


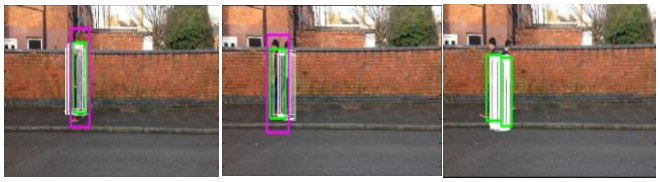
Fig. 14. Multiple target tracking performance. ***Green*** (bounding box): normal tracker and M.I.T; ***Cyan***: similar targets with overlap; ***Purple***: partial occluded target; ***Pink***: full occluded target

The video highlights scenarios where our tracker performs both partial and full occlusion reasoning,

[2] Dataset is available at: http://postgrad.eee.bham.ac.uk/xiaoj/Publications.htm

including detection of full occlusion case followed by target identification after its re-appearance. Similar reasoning is shown in Video 3 where the distinct size of the objects to be tracked did not influence the quality of the tracking results. This video is used later on for comparison purposes.

To better understand the scenario of the target re-appearance, Fig. 15 highlights specific frames of Video 2 showing the target estimation in terms of the their bounding boxes as well as the distribution of the particles with higher weight, beyond a threshold 0.7. Similarly TAB. VIII summarizes the likelihood value associated to each target and number of particles whose weights are beyond the threshold. Especially, out of total 100 particles, there are 16, 40 and 13 particles of occluded target in frame 89, 104 and 108, respectively.



Frame 89 Frame 104 Frame 108

Fig. 15. Bounding box. ***Red***: fully occluded target; ***Green***: non-occluded target; ***White***: particles of occluded target whose likelihood exceed the threshold.

TABLE VIII. LIKELIHOOD OF TARGET ESTIMATION

Frame #	Non-occluded target	Occluded target
89	0.9028	0.7596
104	0.9048	0.7889
108	0.9096	0.8581

Notice that the occluded target in frame 108 has a likelihood value greater than threshold T_r (0.8), which confirmed the target re-appearance and ended the fully occlusion situation.

Next, we compared our collaborative tracking methodology to three other state-of-art trackers. The first one uses our proposed multiple individual tracker (M.I.T) without occlusion reasoning. The second employs the Linear Trajectory Avoidance (LTA) method proposed in [31], which has been proven to provide good accuracy results in case of multiple object tracking. A third approach consists of Struck method [32] due to its superior performances in case of single target tracking (see TAB. IV where it ranked second in overall). In order to monitor the performance of the trackers at various frames, we quantify the overlap at each frame for each target and video sequence. The results are reported in Fig. 16, 18 and 20 where target A corresponds to the left target in the video sequence and target B to the right one. In Video 1 (results shown in Fig.16), one notices, for instance, that M.I.T tracker looses both targets A and B for some frames (when overlap value is close or equal to zero). This is because, for those frames, the tracker does track the same target due to inappropriate occlusion handing mechanism. For LTA tracker, even if it appears to be much better than M.I.T tracker, but still it has also shown slight and brief loose of target A in frame 260. Struck approach also losses target B at several frames. While our robust approach successfully tracks both targets across all frames. Overall results across all frames in terms of MOTA, FNR, FPR and MMR are provided in TAB.IX. Strictly speaking, given that two targets have similar appearance models, so we may likely expect M.I.T to fail when the targets get sufficiently close to each other, while a sort of collision prevention model is applied in LTA to separate them. Our approach not only includes a mechanism for such collision prevention model but also provides a trajectory monitoring procedure which proved to enhance the outcome.

In order to analyze the origin of failure, we represented in Fig. 17 selected scenarios of failure of other trackers. We show some failure cases of other trackers in Fig.17 where the number under the parenthesis is the frame number in Video 1.

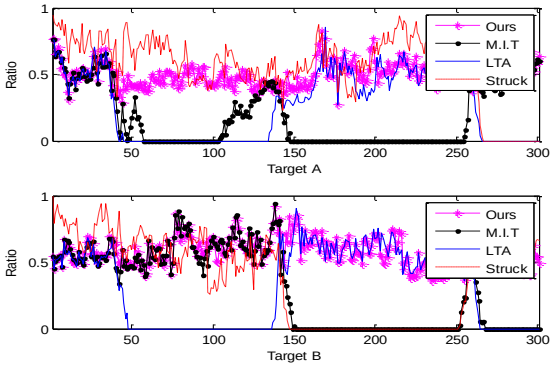
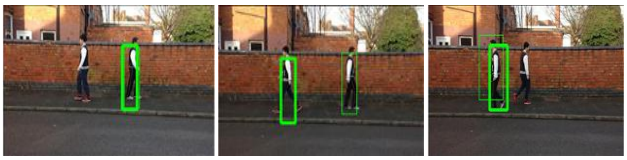


Fig. 16. Video 1: Overlap performance

TABLE IX. OVERALL PERFORMANCE FOR VIDEO 1

	MOTA	FNR	FPR	MMR
Ours	1.0000	0.000	0.0000	0.0000
MIT	0.5183	0.0781	0.0000	0.4037
LTA [31]	0.5781	0.0000	0.0000	0.4219
Struck[32]	0.7674	0.0000	0.0000	0.2326

Fig. 17 reveals that both M.I.T and Struck trackers detected only one single target, while LTA tracker encountered an id-switch.



M.I.T (276) LTA (59) Struck (151)

Fig.17 Example of failures of MIT, LTA and Struck trackers

Results pertaining to Video 2, where the objects to be tracked have distinct appearance but frequent interactions, are highlighted in Fig. 18.

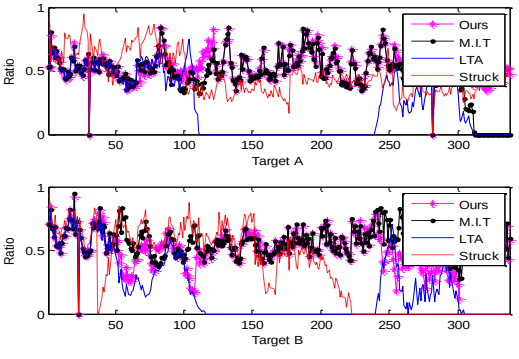


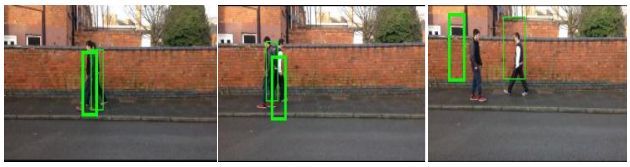
Fig. 18 Video 2: Overlap performance

In this case, M.I.T tracker performs quite well keeping a good separation between target estimates that prevented full occlusion occurrence, and provides performance close to our algorithm. While LTA fails to handle long term target interaction, which yields full occlusions, and thereby, target loss. TAB.X provides global performance results in terms of MOTA, FNR, FPR and MMR metrics. Again the results show that our tracker substantially outperforms the other two trackers.

TABLE X. OVERALL PERFORMANCE FOR VIDEO 2

	MOTA	FNR	FPR	MMR
Ours	0.9956	0.0030	0.0000	0.0015
MIT	0.9571	0.0030	0.0000	0.0399
LTA [31]	0.5281	0.0030	0.0000	0.4689
Struck [32]	0.8240	0.1746	0.0000	0.0015

The failure cases of three alternative trackers in Fig.19 show that the M.I.T performs fairly well when the targets are distinguished from each other, LTA tracker suffers from id-switch. The struck tracker is on the other hand strongly affected by inaccurate estimation, which causes target loss.



M.I.T.(75) LTA (108) Struck (228)

Fig.19 Failures of other trackers in Video 2

Results pertaining to Video 3 are plotted in Fig. 20. In this case, one notices that given the large discrepancy of size of the two objects, both M.I.T and LTA lose target B during the long time occlusion. This is because the size of the occluder is much bigger than the occluded target, making the handling of the long term occlusions inappropriate. Improvements observed when using our tracker in this respect are mainly due to the efficiency of our collaborative tracking to tackle target occlusion. Besides, dark illumination of the sequences together with the use of Haar like feature render the estimation using Struck tracker biased and caused target loss as well. It is also expected that the quality of the training phase in both LTA and Struck negatively contributed to failure of the trackers. Global results related to Video 3 are highlighted in TAB.XI.

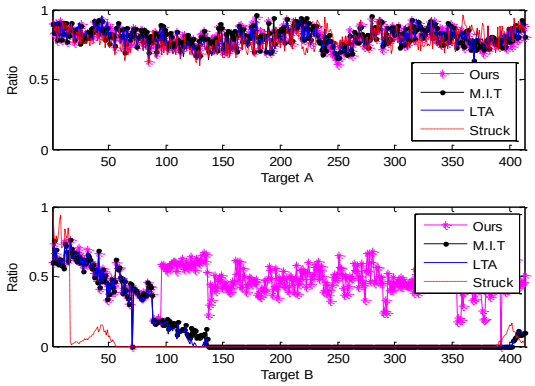


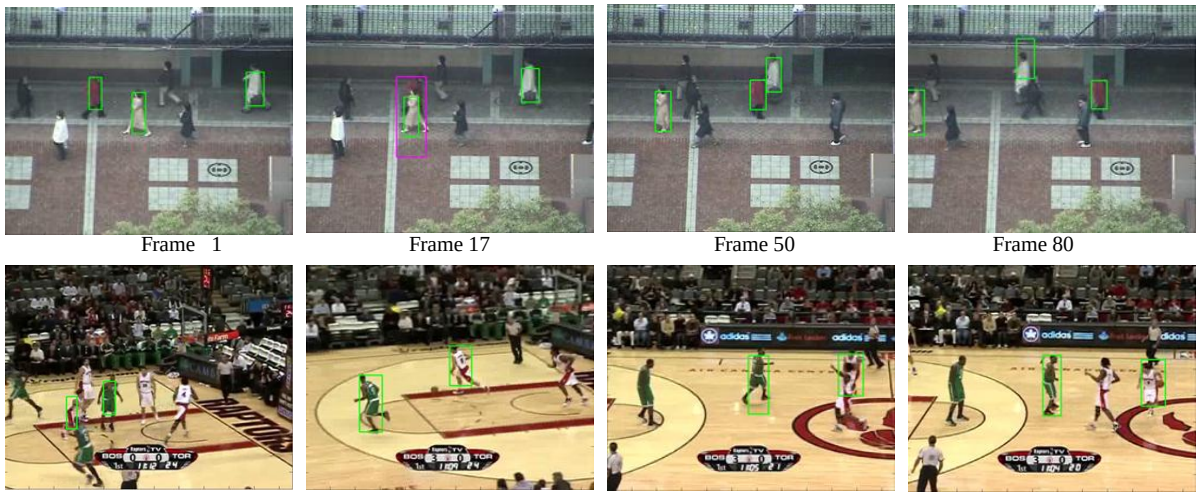
Fig. 20 Video 3: Overlap performance

TABLE XI. OVERALL PERFORMANCE FOR VIDEO 3

	MOTA	FNR	FPR	MMR
Ours	0.9976	0.0024	0.0000	0.0000
MIT	0.6768	0.2215	0.0000	0.1017
LTA [31]	0.6671	0.2470	0.0000	0.0860
Struck [32]	0.5969	0.2954	0.0000	0.1077

B. Benchmark dataset

To better demonstrate the performance of our algorithm, we select two representative sequences, *subway* [35] and *basketball* [36], for our experiment. Fig. 21 illustrates the performance of our tracker on selected frames yielding possible occlusion on both subway and basketball sequences.



Frame 1 Frame 17 Frame 50 Frame 80

Frame 1 Frame 100 Frame 210 Frames 230

Fig. 21. Experiment in benchmark dataset (1st row: Subway; 2nd row: Basketball)

The graph illustrates in case of subway sequence how our algorithm successfully tracked three distinct persons, two of which have close appearance models and with possibility of occurrence of occlusion. While in basketball sequence, two players have been successfully tracked. Notice, that the performance of the tracker may get slightly degraded because of non-homogenous movement of the frames as opposite to the first video sequence because of the (possible) abrupt acceleration of

the players, which makes the basketball sequence more challenging.

In order to compare the overall performance of the trackers across all the frames, the trade-off curves between the MOTA and overlap threshold for subway and basketball sequences are shown in Fig.22 and 23, respectively.

The results confirm the superiority of our tracker as compared to standard M.I.T and LTA trackers for both video sequences. It also shows that LTA approach quite underperforms both other trackers in case of subway video sequence because of long term interactions among the objects, which reinforces the results obtained with our home-build video sequences. On the other hand, M.I.T and LTA provide close performance results in case of fast moving targets of basketball sequence. Especially, both trackers induce situations where the target is lost because of inefficiency of collision adaption in case of LTA and gradual drift towards background of M.I.T tracker because of frequent similar target interaction.

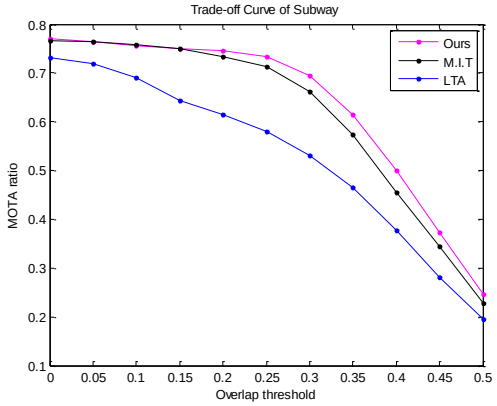


Fig. 22 Trade-off curve of subway

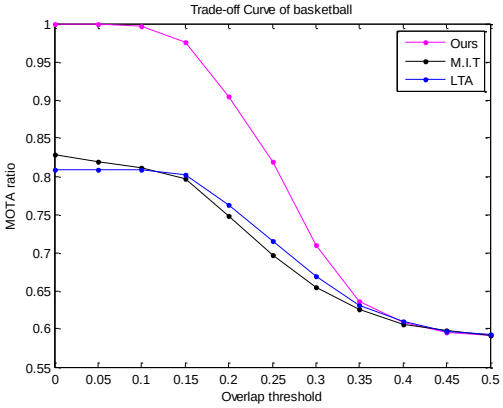


Fig. 23 Trade-off curve of basketball

V. CONCLUSION

In this work, a new multiple target tracking algorithm for visual objects is investigated. The proposal builds on the colour-based particle filter algorithm that was extended in several directions. First, in order to deal with uncertainty arising from background clutter and illumination change, the contextual information is taken into account by enlarging the boundary of the estimated target region, and comparing this with the current observation. Second the distribution of the particles is taken into account through the introduced recursive estimation that restricts the effect of outliers on global estimate of the target. Third, in order to extent the proposal to track multiple objects, although the intuitive use of multiple independent trackers (M.I.T), where each (robust) tracker is associated to an individual target, seems rationale, cautious is required to avoid the problem of occlusion or identity switch. In order to deal with this problem, the distance between the trackers is monitored. For this purpose, a collision prevention model, which prevents tracker *jump-over* scenarios, is introduced, where the appearance similarity scores are employed. In case of (partial) occlusion, a subspace-based method is employed where each particle is partitioned into equal partition, and only the visible parts of the partition are used for tracking. In case of full occlusion, the essence is to reinitialize the particles around the occluder to capture the reappearance of the target. Besides, the tracking algorithm also distinguishes the case where the appearance models do not discriminate between the various targets. In such case, we rather rely on monitoring targets' trajectories. Comparisons with state of art trackers using both home built and open dataset demonstrated the feasibility and the superiority of our proposed tracker to deal with occlusions, clutter and abrupt illumination change. As perspective work, we intend to investigate in more detail the convergence properties of the newly elaborated tracker where more theoretical results are expected.

Acknowledgment:

Miss Jingjing Xiao is supported by Chinese CSC foundation, which is gratefully acknowledged. We also want to thank for the kind help from Jiandi Wei for videos recording. National Science Foundation of PR China under grant number 61171136 is also acknowledged for their financial support

REFERENCES

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, 34, pp. 3-19, 2013.

[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review", *Neurocomputing*, 74(18), pp. 3823–3831, 2011.

[3] Y. Wu, J. Lim and M.H. Yang. "Online Object Tracking: A Benchmark", In: *CVPR*, pp.2411-2418, 2013.

[4]D. Comaniciu, V. Ramesh, and P. Meer. "Kernel-Based Object Tracking", In: *PAMI*, 25(5), pp.564–577, 2003.

[5] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*, pp.886-893, 2005.

[6] O. Tuzel, F. Porikli, and P. Meer. "Region Covariance: A Fast Descriptor for Detection and Classification". In: *ECCV*, pp. 589--600, 2006.

[7] P. Viola and M. J. Jones. "Robust Real-Time Face Detection". In: *IJCV*, 57(2),pp.137–154, 2004.

[8] N. Alt, S. Hinterstoisser, and N. Navab. "Rapid Selection of ReliableTemplates for Visual Tracking". In: *CVPR*, pp. 1355 – 1362, 2010.

[9] G. D. Hager and P. N. Belhumeur. "Efficient Region Tracking With Parametric Models of Geometry and Illumination", *PAMI*, 20(10), pp.1025–1039, 1998.
[10] J. Xiao, R. Stolkin, and A. Leonardis, "An enhanced adaptive coupled-layer LGTracker++", In: *ICCVW*, pp. 137-144, 2013.
[11] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. "Incremental Learning for Robust Visual Tracking" *IJCV*, 77(1), pp.125–141, 2008.
[12] D. Comaniciu, V. Ramesh, and P. Meer, P. "Real-time tracking of non-rigid objects using mean shift". *CVPR*. 2, pp. 142-149.2000
[13] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive kalman filter," *Journal of Visual Communication and Image Representation*, 17(6), pp. 1190–1208, 2006.
[14] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, 21(1), pp. 99–110, 2003.
[15] V. Papadourakis and A. Argyros. "Multiple objects tracking in the presence of long-term occlusions." *Computer Vision and Image Understanding*, 114(7), pp: 835-846, 2010.
[16] A. Yao, X. Lin, G. Wang, and S. Yu, "A compact association of particle filtering and kernel based object tracking," *Pattern Recognition*, 45(7), pp. 2584–2597, 2012.
[17] R. T. Collins, Y. Liu, and M. Leordeanu. "Online Selection of Discriminative Tracking Features". *PAMI*, 27(10):1631–1643, 2005.
[18] H. Grabner, J. Matas, L. V. Gool, and P. Cattin. "Tracking the Invisible: Learning Where the Object Might be." In: *CVPR*, pp.1285-1292, 2010.
[19] B. Liu, J. Huang, L. Yang and C. Kulikowsk, "Robust Tracking Using Local Sparse Appearance Model and K-selection," In: *CVPR*, pp. 1313-1320, 2011.
[20] K. Okuma, A. Taleghani, N. De Freitas, et al. "A boosted particle filter: Multitarget detection and tracking", In: *ECCV*. pp: 28-39, 2004.
[21] W. Du and J. Piater. "Multi-camera people tracking by collaborative particle filters and principal axis-based integration", In: *ACCV*, pp: 365-374,2007.
[22] H. B. Shitrit, J. Berclaz, F.Fleuret and P. Fua, "Tracking Multiple People under Global Appearance Constraints," In *ICCV*, pp. 137-144, 2011.
[23] H. Pirsiavash, D.Ramanan and C. C. Fowlkes, "Globally-optimal Greedy Algorithms for Tracking a Variable Number of Objects," In: *CVPR*, pp. 1201-1208, 2011.
[24] T. Yu and Y. Wu, "Collaborative tracking of multiple targets." In: *CVPR*, pp.834-841, 2004.
[25] A. Bhattacharyya, "On a Measure of Divergence between Two Multinomial Populations," *Sankhya*, 7, pp. 401-406, 1946.
[26] M. Talha and R. Stolkin, "Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data," *IEEE Sensors Journal*, 14(1), pp. 159–166, 2014.
[27]J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter." In: *CVPR*, pp. 1037–1042, 2005.
[28] M. Kristan, et al. "The visual object tracking vot2013 challenge results." In: *ICCVW*, pp. 98-111, 2013.
[29] K. Bernardin and R. Stiefelhagen. "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *Journal of Image and Video Processing*, 2008.
[30] J. Martınez-del Rincon, C. Orrite, and C. Medrano, "Rao–blackwellised particle filter for color-based tracking" P*attern Recognition Letters*, 32(2), pp. 210–220, 2011.
[31] S. Pellegrini, A. Ess, K. Schindler, L. Van Gool "You'll never walk alone: Modeling social behavior for multi-target tracking", In: *ICCV*, pp: 261-268, 2009.
[32] S. Hare, A. Saffari, and P. HS Torr. "Struck: Structured output tracking with kernels". In: *ICCV*, pp.263–270, 2011
[33] X. Mei and H. Ling. "Robust visual tracking using l1 minimization". In: *ICCV*, pp.1436–1443, 2009.
[34] J. F Henriques, R. Caseiro, P. Martins, and J. Batista. "Exploiting the circulant structure of tracking-by-detection with kernels". In: *ECCV*, pp. 702–715, 2012.
[35] F. Pernici and A. Del Bimbo. "Object tracking by oversampling local features," *PAMI*, pp. 2538 – 2551, 2014.
[36] Video benchmark dataset, available: http://cvlab.hanyang.ac.kr/tracker_benchmark/seq/Subway.zip
[37] W. Qu et al., "Real-Time Distributed Multi-Object Tracking Using Multiple Interactive Trackers and a Magnetic-Inertia Potential Model," *IEEE Transactions on Multimedia*, pp. vol. 9, 511-519, 2007.

Jingjing Xiao received her Bachelor and Master degree from College of Mechatronics Engineering and Automation, National University of Defence Technology, China, in 2010 and 2012, respectively. Currently, she is studying in the University of Birmingham, U.K., as a PhD student. Her research interests include computer vision and machine learning.

Dr. Mourad Oussalah holds a PhD and Msc degrees from University of Paris XII in 1998 and 1994. After a postdoctoral experience in KU Leuven Belgium and City University of London, he is since 2003 with University of Birmingham. He is active researcher in Data mining, computer vision and information fusion where he published more than 200 technical papers. He is a Fellow of Royal Statistical Society and executive member of IEEE SMS UK Chapter.