Distractor-Supported Single Target Tracking in Extremely Cluttered Scenes

Jingjing Xiao[1](✉), Linbo Qiao[2], Rustam Stolkin[1], and Aleš Leonardis[1]

[1] University of Birmingham, Birmingham B15 2TT, UK
shine636363@sina.com, {r.stolkin,a.leonardis}@cs.bham.ac.uk
[2] College of Computer, National University of Defense Technology,
Changsha 410073, China
qiao.linbo@nudt.edu.cn

Abstract. This paper presents a novel method for single target tracking in RGB images under conditions of extreme clutter and camouflage, including frequent occlusions by objects with similar appearance as the target. In contrast to conventional single target trackers, which only maintain the estimated target status, we propose a multi-level clustering-based robust estimation for online detection and learning of multiple target-like regions, called *distractors*, when they appear near to the true target. To distinguish the target from these distractors, we exploit a global dynamic constraint (derived from the target and the distractors) in a feedback loop to improve single target tracking performance in situations where the target is camouflaged in highly cluttered scenes. Our proposed method successfully prevents the estimated target location from erroneously jumping to a distractor during occlusion or extreme camouflage interactions. To gain an insightful understanding of the evaluated trackers, we have augmented publicly available benchmark videos, by proposing a new set of clutter and camouflage sub-attributes, and annotating these sub-attributes for all frames in all sequences. Using this dataset, we first evaluate the effect of each key component of the tracker on the overall performance. Then, the proposed tracker is compared to other highly ranked single target tracking algorithms in the literature. The experimental results show that applying the proposed global dynamic constraint in a feedback loop can improve single target tracker performance, and demonstrate that the overall algorithm significantly outperforms other state-of-the-art single target trackers in highly cluttered scenes.

1 Introduction

Visual object tracking remains an open and active research area, despite publication of numerous tracking algorithms over the last three to four decades [1,2]. A particular difficult problem is how to track a target which moves through scenes featuring several other very similar objects, or which moves past extremely cluttered or camouflaged image regions, Fig. 1A. In these situations, it is difficult to distinguish the target using only its appearance information [3]. Additional information, such as dynamic models of the target and nearby distracting image

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 121–136, 2016.
DOI: 10.1007/978-3-319-46493-0_8

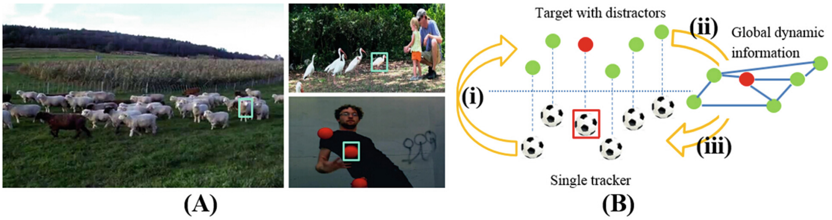

Fig. 1. (A) Sequences with many distractor objects used in our experiments. The cyan bounding boxes depict the single targets we want to track. (B) The proposed tracking framework. The yellow arrows show the feedback loop between the single target information and global dynamic information extracted from the tracker and the distractors. The red bounding box/dot represents the single target we want to track while the green dots denote the distractors. The proposed algorithm: (i) simultaneously tracks the target and the distractors using the proposed robust estimation method; (ii) extracts a global dynamic model from the relative target and distractor trajectories; (iii) feeds the global dynamic information back to the single target tracker to help identify the true target and infer occlusion situations. (Color figure online)

regions, may be useful to support robust tracking [4,5]. Therefore, in this paper, we show how a single target tracker can be used to detect and exploit contextual information. This contextual information is then fed back to the tracker to improve its robustness in problems of tracking a single target which is camouflaged against scenes containing a large number of similar non-target entities, which we call *distractors*.

Our proposed tracker is a *single*-target tracker, in the sense that it is initialised only with a bounding box of a single target in the first frame. However, unlike most single target trackers, it encodes information about other objects or image regions (distractors) with similar appearance to the target, and exploits a global dynamics constraint in a feedback loop to help disambiguate the target from these distractors, as illustrated in Fig. 1B. In scenes with clutter and camouflage, the proposed method detects multiple target-like regions and explicitly models this global information to improve the performance of the single target tracker, using methods which are somewhat analogous to the data association approaches used in multi-target tracking. However, in contrast to multi-target trackers, our proposed method (i) aims at using global information to improve a single target tracker at each frame; (ii) does not assign individual IDs to multiple other objects. The relationship of our tracker to single-target and multi-target trackers is illustrated in Fig. 2.

The main contributions of this paper are: (i) a novel coarse-to-fine multi-level clustering based robust estimation method for online detection and localisation of candidate image regions containing the true target and/or distractors; (ii) a novel global dynamics constraint applied in a feedback loop, which enables the motion of the target and an arbitrary number of distractors to be robustly disambiguated, while also making inferences about occlusion situations; (iii) for

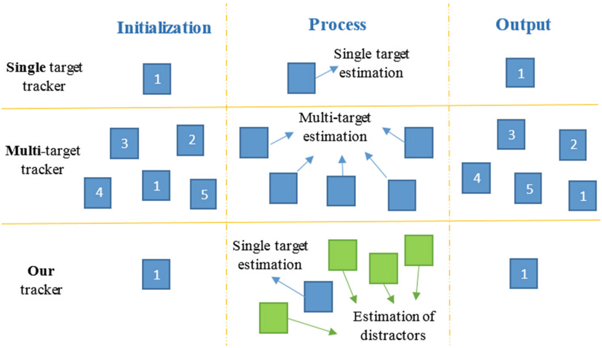

Fig. 2. Relationships between single target trackers, multi-target trackers and our proposed tracker during tracking. Single target trackers do not explicitly model information about distractor objects. Multi-target trackers model and identify *multiple* target regions, exploiting additional initialised prior knowledge. In contrast, our tracker is initialised in the same way as single-target trackers, but automatically detects and learns models for multiple distractor entities on the fly, to help improve the performance of the *single* target tracker.

performance evaluation, we propose a new set of sub-attributes to describe different kinds of cluttered scenes, and we augment publicly available benchmark data by per-frame annotations of all sequences with all sub-attributes. We perform two sets of experiments using publicly available ground-truthed datasets. First, on highly cluttered scenes, we (i) compare our tracker against other state-of-the-art single target trackers, demonstrating superior performance of the proposed algorithm, and (ii) study our tracker by evaluating the effectiveness of each designed component. Secondly, for an overall assessment of the tracker, we also evaluate its tracking performance on non-cluttered scenes from OTB100 [1] dataset, again with favourable results. The remainder of this paper is organised as follows. We review related work in Sect. 2. Section 3 explains technical details of our proposed tracker. Section 4 presents and discusses experimental results. Section 5 provides concluding remarks.

2 Related Work

We first review works on *single*-target tracking in highly cluttered scenes. Then, we review some related data association methods used in *multi*-target tracking literature to make a distinction between that work and our own (a single-target tracker which additionally models the global dynamics of multiple distractors).

To track a single target robustly in the presence of clutter, the tracker should learn and exploit contextual information. In [6,7] it was observed that non-target objects, known as *supporters*, may sometimes be associated with a target and can be used to help infer its position. However, in highly cluttered scenes, e.g. Fig. 1A, it may not be possible to find supporters that persistently move around the target with strong motion correlation. In contrast, we notice that identifying *distractors* in contextual clutter can also help robustify target identification. Other work [8–10] detected distractor regions with similar appearance to the target. However, tracking accuracy of such methods heavily depends on the pre-defined spatial density of samples, where sparse sampling cannot adequately distinguish adjacent objects, and dense sampling is computationally expensive. Even with dense sampling, such methods can still fail to distinguish adjacent or overlapping objects. In contrast, we propose a robust estimation method which uses a multi-level clustering scheme to efficiently search for objects at progressively finer granularities, and distinguishes inter-occluding objects using a novel method based on the disparity between mean and mode samples. Note that methods such as [8–10] maintain multiple image regions as target candidates, but do not specifically decide which region is the target at each frame. In contrast, our proposed method detects and learns distractors on-the-fly, and then exploits global target-distractor dynamics constraints to enable deterministic identification of the target at each frame. Appearance matching scores were used to distinguish the target from distractors in [11]. However, such algorithms are prone to failures when the target is camouflaged, occluded or undergoing deformation, when appearance matching methods can cause the tracker to erroneously fixate on clutter. A unified model to select the best matching metric (attribution selection) and most stable sub-region of the target (spatial selection) for tracking was proposed in [12]. Hong et al. [13] learned a discriminative (matching) metric that adaptively computed the importance of different features, and online adaptive attribute weighting was also proposed in [14–16]. Posseger et al. [17] recently proposed a distractor-aware target model to select salient colours in single target tracking. However, none of the methods [12–17] actively searches and memorises the trajectories of the distractors in scenes, or exploits a global dynamic constraint to improve single target tracking. In addition, our paper addresses video sequences that are so extreme that both target and distractors may have *identical* appearance, and cannot be disambiguated by any appearance features.

We now discuss the conceptual differences between our proposed global dynamic constraint and data association methods used in multi-target trackers, as illustrated in Fig. 2. Berclaz et al. [18] reformulated the data association problem as a constrained flow optimization convex problem, solved using a k-shortest paths algorithm. However, the computational cost of generating k paths is quite high, especially for our problem of finding a single target at each frame. Moreover, this method first obtains detections for every frame throughout an entire video sequence, and then mutually optimises the target IDs over all frames. Such post-processing methods cannot be used for online target tracking. Shitrit et al. [19] also relaxed the data association problem as a convex optimization problem which explicitly exploited image appearance cues to prevent erroneous identity switch. However, appearance cues are not sufficiently discriminating to distinguish between the target and the distractors in extremely challenging videos which we tackle in this paper, as shown in Fig. 1. Dicle et al. [3] utilized motion dynamics to distinguish targets with similar appearance, in order to reduce instances of target mislabelling and recover missing (occluded) data. However, their algorithm requires the number of targets to be known a-priori. In contrast, our method handles situations where the number of distractors is unknown and has to be learned on the fly during tracking. Chen et al. [20] proposed a constrained sequential labelling to solve the multi-target data association problem, which utilized learned cost functions and constraint propagation from captured complex dependencies. However, their approach is only designed to handle the case of piece-wise linear motion. The single target tracker of [21] was extended to online multi-target tracking [22] by using global data association. However, candidate target regions must be densely sampled which can be extremely computationally expensive. Moreover, the global identity-aware network flow graph of [22] depends heavily on target appearance models, which have difficulty in handling highly cluttered scenes, especially when both target and distractors share identical appearance. In [23], we learned and exploited the global movement of sports players to inform strong motion priors for key individual players. Information from the global team-level context dynamics enabled the tracker to overcome severe situations such as inter-player occlusions. However, the proposed context-conditioned latent behavior models do not readily generalise to non-sports tracking situations.

In contrast to the above-mentioned works, our proposed tracker explicitly exploits contextual information to detect and learn nearby distractors on-the-fly. It then simultaneously builds a tracking memory of both the target and the distractors, which is used to compute an online-learned global dynamic constraint which is finally fed back to help robustify the single-target tracker.

3 Proposed Distractor-Supported Single-Target Tracking Method

The proposed method consists of two steps. The first step uses the proposed robust estimation with coarse-to-fine multi-level clustering to find candidate image regions for the target and any distractors. The second step distinguishes the target from the distractors, and infers occlusion situations, by feeding back the extracted global dynamic constraint (based on the motion history of both the tracker and the distractors) to the single target tracker.

3.1 Robust Estimation with Coarse-to-fine Multi-level Clustering

Our proposed tracking algorithm first propagates a set of samples drawn from the region around the target position estimated at the preceding frame. We then propose a multi-level clustering-based robust estimation method to find regions that are similar to the target in the new frame. Multiple feature modalities and spatial information are used, level by level, in a coarse-to-fine sampling manner to incrementally achieve better results, shown in Fig. 3. This approach resolves the tradeoff between robustness and tracking speed, by first performing sparse

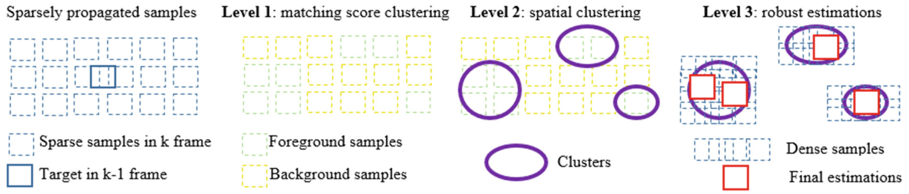
Sparsely propagated samples
Level 1: matching score clustering
Level 2: spatial clustering
Level 3: robust estimations
Sparse samples in k frame
Target in k-1 frame
Foreground samples
Background samples
Clusters
Dense samples
Final estimations

Fig. 3. Coarse-to-fine multi-level clustering-based robust estimation is used to find image regions containing the target or distractors. The algorithm: (i) sparsely propagates samples around the target position from the previous frame; (ii) clusters the propagated samples into two groups (foreground/background samples) according to their associated matching scores; (iii) clusters the foreground samples according to their spatial distribution; (iv) densely samples each clustered foreground region to perform robust estimations.

sampling to find initial candidates, and later applying dense sampling to a small subset of image regions where needed.

The algorithm begins by propagating only a sparse set of samples, with colour features initially used to compute matching scores for each sample. First, clustering is carried out according to colour matching scores to classify samples into *foreground* and *background* sets (level 1 clustering), defined as those with high and low matching scores respectively. Next, the spatial distribution of level 1 foreground samples is used to sub-cluster neighbouring samples (level 2 subclustering). For each level 2 cluster, we then apply a dense sampling, using an additional feature (HOG) for robust estimation (level 3 cluster subdividing). Note that we use the term *foreground* in a special sense, to denote *both the target and the distractors*. Everything else is called *background*.

A. Level 1 Clustering. The algorithm samples a sparse set of N_p locations surrounding the target location in a uniform way. The positions of the samples at the kth frame are denoted by $\{\mathbf{p}_k^i\}_{i=1,\ldots,N_p}$. As colour histograms are acknowledged for their simplicity, computational efficiency, invariance to scale and resolution change [24], we first extract a colour histogram from each sample and compare it to the target appearance model to get matching scores $\{w_{C,k}^i\}_{i=1,\ldots,N_p}$, where C indicates the colour feature. Within the information of the sample distributions and their associated matching scores, we use $\mathbf{x}_k^i = \{\mathbf{p}_k^i, w_{C,k}^i\}$ as the feature vector for a Gaussian Mixture Model in order to cluster the samples into two groups: foreground samples and background samples, according to Eq. 1:

$$p(\mathbf{x}_k^i;\theta) = \sum_{\mathbb{C}=1}^{2} \alpha_{\mathbb{C}} \mathcal{N}(\mathbf{x}_k^i; \mu_{\mathbb{C}}, \sum\nolimits_{\mathbb{C}}) \tag{1}$$

where $\alpha_{\mathbb{C}}$ is the weight of the cluster $\mathbb{C}$, $0 < \alpha_{\mathbb{C}} < 1$ for all components, and $\sum_{\mathbb{C}=1}^{2} \alpha_{\mathbb{C}} = 1$, where μ and $\sum$ are the mean and variance of the corresponding cluster. The parameter list:

$$\theta = \{\alpha_1, \mu_1, \sum\nolimits_1, \alpha_2, \mu_2, \sum\nolimits_2\} \tag{2}$$

defines a Gaussian mixture model, which is estimated by maximising the likelihood [25]. The mean matching score of samples in each cluster is denoted by $\bar{w}_{C,k}^{\mathbb{C}}$. Then, all samples in the cluster with the highest mean score are regarded as foreground samples, Eq. 3:

$$\mathbb{R}_f(\hat{i}) = 1, \quad \text{if} \quad \hat{i} = \arg\max \bar{w}_{C,k}^{\mathbb{C}(i)} \tag{3}$$

where $\mathbb{R}_f$ denotes whether sample $\hat{j}$ is regarded as foreground. The selected foreground samples $\mathbf{p}_k^i$ will next be used for level 2 sub-clustering using additional features, while all samples in the cluster with lower mean score are regarded as background and are discarded.

B. Level 2 Sub-clustering. In a highly cluttered environment, there may be many false positives among those samples labeled as foreground, caused by distractors (non-target image regions with target-like appearance). To distinguish individual objects in the scene (the target and the distractors) we therefore sub-cluster the samples within all level 1 clusters according to their spatial distribution:

$$\mathbb{C}_{sub}(i,j) = 1, \quad \text{if} \quad \mathbb{N}(i,j) = 1, \quad i,j \in \mathbb{R}_f \tag{4}$$

where $\mathbb{N}(i,j)$ denotes whether samples i and j are neighbours. $\mathbb{C}_{sub}(i,j) = 1$ labels samples i and j as belonging to the same sub-cluster. $\mathbb{R}_f$ represents the foreground sample cluster (level 1 cluster). Noticeably, the performance of this spatial distribution-based clustering method depends on the spatial density of propagated samples. If the samples are sparsely distributed, it is likely that a level 2 sub-cluster may contain more than one object (a similar problem was identified in [9,10]).

Figure 4 illustrates the results after level 1 and level 2 clustering, using a frame from the *Juggling* sequence. Even if there is a gap between two adjacent objects (Fig. 4A), it can be difficult to distinguish them using a sparse sampling density, Fig. 4B. Therefore, we next proceed to another level (level 3 cluster

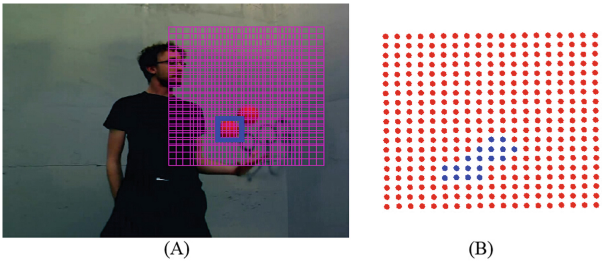

Fig. 4. Failure mode of levels 1 and 2 clustering, due to sparse sampling. (A) red grid denotes the sparsely distributed samples. Blue rectangle is the estimated object; (B) red dots denote background samples while blue dots denote foreground samples which have been erroneously merged into a single foreground cluster. Clearly levels 1 and 2 clustering can fail to disambiguate two adjacent objects. (Color figure online)

subdividing), where the foreground regions identified by levels 1 and 2 are more densely sampled and an additional appearance feature is added to achieve finer scale disambiguation.

C. Level 3 Robust Estimation with Cluster Subdividing. After we obtain the set of foreground samples from levels 1 and 2 clustering, we densely sample the region inside each level 2 sub-cluster to further improve the localisation of target and distractor regions. Each level 2 sub-cluster was obtained using colour features for matching, and all level 2 foreground samples therefore already have a high colour matching score. Therefore an additional feature modality is needed to achieve further disambiguation of target and distractor regions. At level 3 the algorithm therefore applies HOG features to compute the matching scores of the new samples, using a kernelised correlation filter [26].

Within each densely re-sampled level 2 sub-cluster, the most straightforward way to identify the object region is to search for the sample with the highest HOG feature matching score. However, as shown in Fig. 4, sometimes a coarse level 2 cluster may contain more than one object. If the target undergoes deformation, then a distractor within the same cluster often triggers a high matching score. In [24,27] they tried to detect the target by applying the expectation operator over the distributed samples with associated weights (i.e. taking the likelihood-weighted mean of all samples). However, the expectation estimation might be highly erroneous when multiple similar objects are present in the scene [28]. For example, taking the mean location of two similar objects will give an estimated location which lies on a background region, midway between both samples. To overcome this problem, we observe that the spatial ambiguity between the sample with the highest matching score (the *mode*) and the location of the *mean* sample (derived from the expectation operator) can indicate potential distractions within a cluster, and enable robust estimation.

Within the dense level 3 samples, the initial estimate of the object inside each cluster is taken to be the sample with the highest HOG matching score (i.e., the mode sample), denoted by $\mathbf{p}_k^{\mathbb{C}_{sub}(i_h)}$. We also use the expectation operator over all samples in the cluster to compute the mean sample:

$$\bar{\mathbf{p}}_k^{\mathbb{C}_{sub}} = \sum_{i=1}^{N_{\mathbb{C}}} w_{H,k}^{\mathbb{C}_{sub}(i)} \mathbf{p}_k^{\mathbb{C}_{sub}(i)} \tag{5}$$

where $w_{H,k}^{\mathbb{C}_{sub}(i)}$ is the associated HOG feature matching score of the dense sample i inside level 2 sub-cluster $\mathbb{C}_{sub}$ and $N_{\mathbb{C}}$ is the number of samples inside each cluster. If the overlap between $\mathbf{p}_k^{\mathbb{C}_{sub}(i_h)}$ and $\bar{\mathbf{p}}_k^{\mathbb{C}_{sub}}$ is small, it suggests there is another distractor inside the cluster, which is on the opposite side of $\mathbf{p}_k^{i_h}$ compared to $\bar{\mathbf{p}}_k$, see Fig. 5.

If we denote foreground samples in the other half of the cluster as $\mathbb{R}_{f,\mathbb{C}_{sub}/2}$, then a second object's location is estimated by:

$$\begin{aligned} \hat{i} &= \arg\max w_{H,k}^{\mathbb{C}_{sub}(i)} \\ \textbf{s.t.} \quad i &\in \mathbb{R}_{f,\mathbb{C}_{sub}/2}, \quad \mathbf{p}_k^{i_h} \cap \bar{\mathbf{p}}_k^{\mathbb{C}_{sub}} < \zeta \end{aligned} \tag{6}$$

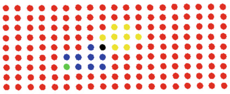

Fig. 5. Sub-image of Fig. 4B with level 3 clustering. Green dot denotes mode sample $\mathbf{p}_k^{\mathbb{C}_{sub}(i_h)}$. Black dot is the mean sample $\bar{\mathbf{p}}_k^{\mathbb{C}_{sub}}$. Yellow dots denote foreground samples in the other half of the same cluster. (Color figure online)

where ζ is the overlap threshold. This method will iteratively estimate the potential distractors inside each cluster until $\mathbf{p}_k^{i_h}$ and $\bar{\mathbf{p}}_k$ have significant overlap. Note that the difference between mode sample and mean sample is utilised in a novel way to indicate the search direction, which helps find the objects quickly, even when partly occluded, as illustrated in Fig. 7.

The final estimations from all clusters indicate "foreground" regions that might contain either the target or distractors, denoted by $\{\mathbf{p}_{o,k}^i\}_{i=1...N_{o,k}}$ where $N_{o,k}$ is the number of observed foreground regions (note we use foreground to refer to both the target and target-like distractors).

So far, we have presented a multi-level clustering-based robust estimation method with coarse-to-fine sampling to detect target-like regions. The method reduces the computational cost compared to dense sampling over the entire image, while improving tracking accuracy compared to methods using a fixed spatial sampling density. The algorithm will next combine the motion history information of both the target and the distractors to build a global dynamic constraint, described in Sect. 3.2. This global information will be fed back to the single-target tracker, deterministically associating a single foreground region to the target and also detecting occlusion situations.

3.2 Global Dynamic Constraint in a Feedback Loop

In highly cluttered scenes, motion cues are important for overcoming the ambiguity caused by appearance similarities between the target and the distractors [3]. Therefore, we use motion history of the target and the distractors to build a global dynamics constraint and feed it back to the individual tracker, which deterministically associates a single foreground region to the true target and prevents the estimated target location from erroneously jumping to a distractor during occlusion or extreme camouflage interactions.

A. Global Motion Regression Model of the Target and the Distractors. During rapid camera motion, the image coordinates of the objects (i.e., of the target and/or distractors) can jump abruptly. However, the *relative* positions between the objects remain relatively stable. Therefore, our global motion model is generated from the relative positions between the tracked target and the surrounding distractors. The multi-level clustering-based robust estimation (described in the previous section) outputs multiple detected foreground objects $\{\mathbf{p}_{o,k}^i\}_{i=1...N_{o,k}}$ at the kth frame. We now re-write the coordinates of these objects as:

$$\mathbf{p}^i_{o,k} = \bar{\mathbf{p}}_{o,k} + \Delta\mathbf{p}^i_{o,k} \tag{7}$$

where $\Delta\mathbf{p}^i_{o,k}$ represents the relative displacement between the ith object location and the spatial distribution centre $\bar{\mathbf{p}}_{o,k} = \frac{1}{N_{o,k}}\sum_{i=1}^{N_{o,k}} \mathbf{p}^i_{o,k}$ of all object position estimates.

Over a short time interval, the underlying dynamics of the target can reasonably be approximated as a linear regression model [3]. The global (relative) motion of the target in frame κ is then predicted by a linear regression model: $\Delta\mathbf{p}_{t,\kappa} = \beta_0 + \beta_1\kappa + \varepsilon_\kappa$, where β_0, β_1 are the coefficients and $\boldsymbol{\varepsilon}_k$ is a noise term. To estimate the parameters, the algorithm minimises the sum of squared residuals $\sum_{i=1}^{k-1} \varepsilon^2_\kappa$, where $\hat{\beta}_0, \hat{\beta}_1$ is obtained from the historic information of the relative position of the target, by least squares estimates. The predicted relative position of the target at frame k is:

$$\Delta\hat{\mathbf{p}}_{t,k} = \hat{\beta}_0 + \hat{\beta}_1 k \tag{8}$$

Note that the relative positions of the target and the distractors implicitly encode global information about the scene dynamics. The relative position of foreground object i at the kth frame can be denoted by $\Delta\mathbf{p}^i_{o,k}$, which is computed from Eq. 7. Note that our tracking algorithm is only concerned with solving the *single* target tracking problem, and does not assign or maintain individual IDs for all foreground objects in the scene. Using the relative target position predicted by the global motion model, we can calculate likelihood of a foreground object being the true target as:

$$w^i_{D,k} = e^{-|\Delta\mathbf{p}^i_{o,k} - \Delta\hat{\mathbf{p}}_{t,k}|} \tag{9}$$

where $w^i_{D,k}$ denotes the dynamic similarity score between the predicted target relative position $\Delta\hat{\mathbf{p}}_{t,k}$ and the relative position $\Delta\mathbf{p}^i_{o,k}$ of the ith foreground object.

Intuitively, the robustness of this dynamic similarity score, in Eq. 9, corresponds to the complexity and stability of the spatial distribution of the the detected foreground objects. If the number of detected foreground objects changes dramatically, this indicates either potential occlusion or newly emerged distractors.

B. Handling Dynamic Numbers of Distractors. While modelling the global dynamics, it is crucial to be able to handle situations where the number of detected foreground objects is changing. In such situations, the relative positions can be highly noisy or even invalid because of newly emerged/disappeared objects.

Newly emerged or disappeared foreground objects might either be the target or the distractors. Therefore, we use the image coordinates to associate each detected object i with a target-like dynamic matching score $w^i_{t,k}$ and distractor-like dynamic matching scores $w^{i,m}_{d,k}$, computed by:

$$\begin{cases} w^i_{t,k} = e^{-|\mathbf{p}^i_{o,k} - \mathbf{p}_{t,k-1}|} \\ w^{i,m}_{d,k} = e^{-|\mathbf{p}^i_{o,k} - \mathbf{p}^m_{d,k-1}|} \end{cases} \tag{10}$$

where $\mathbf{p}_{t,k-1}, \mathbf{p}^{m}_{d,k-1}$ are the positions of the target and the m-th distractor in the $k-1$th frame. $\mathbf{p}^{i}_{o,k}$ is the ith detected object at frame k. Here, the exponential function is applied to normalise the likelihood value to occupy the range $(0, 1)$. The detected object corresponding to the target should have a high target-like dynamic matching score and also a low distractor-like dynamic matching score, giving a *global* dynamic score $w^{i}_{D,k}$ for the ith object as:

$$w^{i}_{D,k} = \begin{cases} N_{d,k-1} w^{i}_{t,k} / \sum_{m=1}^{N_{d,k-1}} w^{i,m}_{d,k} \ , & N_{d,k-1} \neq 0 \\ \qquad w^{i}_{t,k} \qquad , & others \end{cases} \tag{11}$$

where $N_{d,k-1}$ is the number of distractors in the $k-1$th frame.

C. Global Dynamic Constraint Fed Back to the Single-target Tracker. Our proposed algorithm feeds the generated global dynamic information back to the single-target tracker to constrain the estimated target trajectory and detect occlusion situations. Next, the newly estimated target state is used to update the global dynamic information for successive frames, as shown in Fig. 1B. With modelled global information, the final target region is optimally assigned to the candidate region with the highest dynamic similarity score $w^{i}_{D,k}$ by Eq. 12:

$$\begin{aligned} \hat{i} &= \arg\max \ w^{i}_{D,k} \\ \textbf{s.t.} \ \ w^{i}_{t,k} &\geq \lambda_d max \ \{w^{i,m}_{d,k}\}_{m=1,\ldots,N_{d,k-1}} \end{aligned} \tag{12}$$

where $w^{i}_{D,k}$ is computed from Eq. 9 when the number of detected foreground objects is stable, and from Eq. 11 when the number of detected objects is changing. $N_{d,k-1}$ represents the number of the distractors in the $k-1$th frame while λ_d is a scaling factor in the range between 0 and 1. After confirming the target, the appearance model is updated using a linear combination of the reference model and the observation [24,29]. The global dynamic constraint is updated accordingly (Eqs. 9, 11).

4 Experimental Results

We first evaluate our proposed tracker by analysing the contributions of each of the key components (robust estimation with cluster subdividing, and the global dynamic constraint) on overall performance. Next we compare our tracker against the other state-of-the-art trackers which were ranked highest in recent benchmark studies [1,2,30]. Section 4.1 analyses performance specifically on highly cluttered scenes. Section 4.2 tests on all other scenes from OTB100 [1], confirming that our method also performs competitively on uncluttered scenes.

Evaluation Metrics. In this paper, we compare trackers in terms of the area under the curve (AUC) of the overlap rate curve [1]. **Implementation**. The proposed algorithm was implemented in Matlab2014a (linked to some C components) using an Intel Core i5-3570 CPU, giving average speed of 20.23 fps on non-cluttered scenes, and 4.01 fps on highly cluttered scenes with overhead computation cost from global dynamic model. All sequences and the code are publicly available.

4.1 Experiment on Highly Cluttered Dataset

Datasets. We have selected 28 highly cluttered sequences from publicly available data-sets [1–3,22]. Note that we do not use the full datasets in these first tests because: (i) these large datasets only contain a few sequences featuring extreme clutter and camouflage, which this paper specifically addresses; (ii) testing on all sequences introduces confounding factors (non-clutter conditions) making it hard to disambiguate the true capabilities of each algorithm to tackle clutter and camouflage. To gain a deeper understanding of the tracker performance on cluttered scenes, we propose a new set of *sub-attributes* for clutter and camouflage: shape clutter, colour clutter, camera motion-caused camouflage motion, self-moving camouflage. We have *per-frame* annotated all sequences with all these sub-attributes.

A. Evaluation of the Tracker Sub-components. In this section, we decompose the method and evaluate the contribution of each of the key components to the overall performance. In the experiment, the baseline algorithm applies the colour feature used in Sect. 3.1A to estimate the target position from the sample with the highest matching score. Next, we add HOG feature as described in Sect. 3.1C to identify the target region. Since the data association method SMOT [3] is explicitly designed for simultaneously tracking multiple targets which share similar appearance, we use this multi-target tracker to evaluate the effectiveness of the global dynamic constraint. Note that the original SMOT [3] is initialised with ground-truth positions for all objects (potential regions that contain the target or distractors). To conduct a meaningful comparison, we input the same detections from our proposed robust estimation to SMOT for data association and output the optimized path for the target. We provide the AUC results [1] of the decomposed algorithm for single target tracking, tested (i) over the entire dataset and (ii) for the frames corresponding to particular sub-attributes, Table 1.

Table 1 shows that our proposed multi-level clustering-based robust estimation improves tracking performance. The performance of our method versus SMOT [3] demonstrates the effectiveness of our proposed global dynamic

Table 1. AUC for the decomposed single target tracking algorithm tested in extremely cluttered scenes. **B**: baseline algorithm (only colour feature); **H**: HOG feature used in Sect. 3.1C; **GDC**: global dynamic constraint in Sect. 3.2. (red: best performance; blue: second best performance).

Tracker	Overall	Clutter type		Camouflage motion	
		Colour	Shape	Camera-caused motion	Self-motion
B	6.204	5.6072	6.0677	5.7518	6.1285
B+H	7.8030	7.2776	7.6666	6.4885	7.6789
B+H+SMOT [3]	7.3545	6.7536	7.2065	6.5659	7.2558
B+H+GDC	8.7108	8.4086	8.5921	7.0662	8.6388

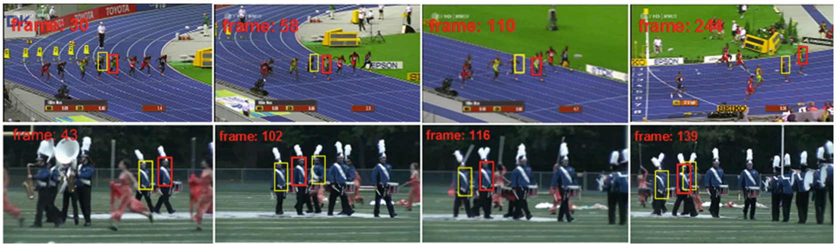

Fig. 6. Performance of our proposed single target tracker in extremely cluttered and camouflaged scenes. First row: *bolt 1*; second row: *marching*. Red bounding box: the target; yellow bounding box: adjacent distractors. (Color figure online)

Table 2. AUC for single target tracking performance in extremely cluttered scenes. Our proposed method significantly outperforms all compared methods on all sub-attributes. (red: best performance; blue: second best performance)

Tracker	Overall	Clutter type		Camouflage motion	
		Colour	Shape	Camera-caused motion	Self-motion
CT [11]	2.7215	2.5956	2.7984	1.9619	2.7641
CPF [24]	5.0872	4.1938	4.9113	4.9852	4.9120
Struck [21]	5.8647	5.2944	6.0627	3.6934	6.0705
SCM [31]	6.4292	5.6997	6.6102	4.1297	6.5985
KCF [29]	7.5602	6.9372	7.5118	5.5188	7.5591
HCF [32]	7.6767	7.0615	7.9109	6.5943	7.9219
Ours	8.7108	8.4086	8.5921	7.0662	8.6388

constraint. Since SMOT algorithm has difficulty handling scenes with highly dynamic number of distractors, it associates the wrong object to the target, impeding tracking performance. Of note, our proposed tracking method runs at 4.01 fps, while SMOT has a speed of 1.86 fps.

Figure 6 illustrates the strong performance of our proposed tracker in extreme clutter and camouflage. Distractors, detected and learned online by our tracker, are indicated by yellow bounding boxes, while the true target is shown with a red bounding box. In frame 139 of sequence *marching*, one distractor shares a major overlap with the target, however our proposed multi-level clustering process can still very accurately disambiguate and localise these two objects.

B. Overall Performance Comparison. To evaluate tracking performance under highly cluttered conditions, our proposed algorithm is compared against several state-of-the-art single target trackers including KCF [29], Struck [21], SCM [31], CPF [24] and the latest CNN-based tracker HCF [32], which were highly ranked in recent benchmark studies [1,2,30]. The CT algorithm [11] is considered as the most closely related work to ours, thus it also takes part in the comparison. We provide the AUC results [1] of each tracker in Table 2. The trade-off overlap rate curve is shown in Fig. 7.

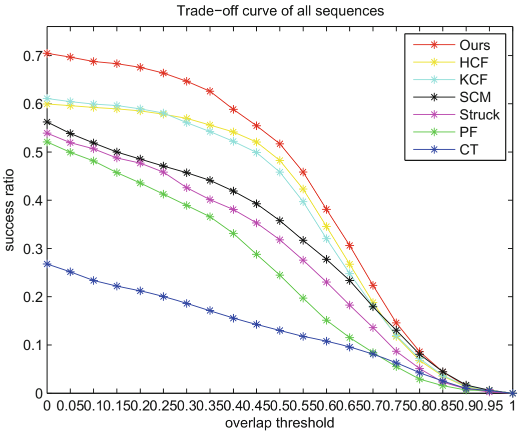

Fig. 7. The trade-off overlap rate curve of single target trackers, tested on 28 videos featuring highly cluttered scenes.

Our proposed tracker outperforms all compared trackers, both overall and also in all sub-attribute categories. KCF [29] and HCF [32] are both based on correlation filters but using different features. Since HCF applies the latest CNN features, it slightly outperforms KCF (using HOG). Note that our proposed method, even without our proposed global dynamic constraint (shown in Table 1), outperforms HCF (Table 2). This is because HCF densely samples the regions around the target, while our coarse-to-fine searching mechanism searches over an initially larger area to progressively finer granularities. CT [11] does exploit contextual information, but the algorithm is still based primarily on appearance matching. Since CT exploits more distracting information which is not properly eliminated, it performs the worst out of the compared methods.

4.2 Experiment on Non-cluttered Dataset

To check how the proposed algorithm performs on non-cluttered scenes, we also tested our algorithm on non-cluttered sequences (94 seq) from OTB100 [1], excluding the already used highly cluttered sequences. The ranks on the non-cluttered scenes are: HCF (11.04, AUC score), *ours* (9.78), KCF (9.44), Struck (9.34), SCM (9.00), CPF (6.82). HCF with CNN-based rich features achieves the best results on the non-cluttered sequences when handling other confounding factors, followed by our tracker with comparably good results.

5 Conclusions

In this paper we presented a novel method for tracking a single target in scenes of extreme clutter and camouflage. In contrast to conventional tracking algorithms which only maintain information about the target, the proposed algorithm incorporates a novel multi-level clustering method for online detection and learning of target-like contextual image regions, called distractors. To disambiguate the target's path among the distractors, a global dynamic constraint is proposed in a feedback loop to improve the single target tracker, and occlusion situations are also detected when no likely target path is found. The proposed method successfully prevents the estimated target location from erroneously jumping to distractors during occlusions or camouflage interactions. To evaluate our tracker, we have introduced a new set of sub-attributes, and have per-frame annotated a number of public benchmark test sequences with these sub-attributes. Using this dataset featuring extreme clutter and camouflage, we have first demonstrated the contribution of each key component of the tracker to the overall tracking performance, and then compared our tracker against highly ranked target tracking algorithms from the literature, demonstrating that our proposed method significantly outperforms other state-of-the-art trackers. In addition, we tested the tracker on non-cluttered scenes, where it also achieves competitive performance.

Acknowledgement. We acknowledge MoD/Dstl and EPSRC for providing the grant to support the UK academics (Ales Leonardis) involvement in a Department of Defense funded MURI project. This work was also supported by EU H2020 RoMaNS 645582 and EPSRC EP/M026477/1.

References

1. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR, pp. 2411–2418. IEEE (2013)
2. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Cehovin, L.: A novel performance evaluation methodology for single-target trackers. PAMI (2015). http://ieeexplore.ieee.org/document/7379002/
3. Dicle, C., Camps, O.I., Sznaier, M.: The way they move: Tracking multiple targets with similar appearance. In: ICCV, pp. 2304–2311. IEEE (2013)
4. Kristan, M., Kovacic, S., Leonardis, A., Pers, J.: A two-stage dynamic model for visual tracking. IEEE Trans. Syst. Man Cybern. **40**(6), 1505–1520 (2010)
5. Xiao, J., Stolkin, R., Leonardis, A.: Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models. In: CVPR (2015)
6. Grabner, H., Matas, J., Van Gool, L., Cattin, P.: Tracking the invisible: Learning where the object might be. In: CVPR, pp. 1285–1292, June 2010
7. Yang, M., Wu, Y., Hua, G.: Context-aware visual tracking. PAMI **31**(7), 1195–1209 (2009)
8. Vermaak, J., Doucet, A., Pérez, P.: Maintaining multimodality through mixture tracking. In: ICCV, pp. 1110–1116. IEEE (2003)
9. Okuma, K., Taleghani, A., Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: multitarget detection and tracking. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24670-1_3
10. Yang, C., Duraiswami, R., Davis, L.: Fast multiple object tracking via a hierarchical particle filter. In: ICCV, vol. 1, pp. 212–219. IEEE (2005)
11. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: CVPR, pp. 1177–1184. IEEE (2011)

12. Jiang, N., Wu, Y.: Unifying spatial and attribute selection for distracter-resilient tracking. In: CVPR, pp. 3502–3509 (2014)
13. Hong, Z., Mei, X., Tao, D.: Dual-force metric learning for robust distracter-resistant tracker. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 513–527. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33718-5_37
14. Talha, M., Stolkin, R.: Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data. IEEE Sens. J. **14**(1), 159–166 (2014)
15. Stolkin, R., Rees, D., Talha, M., Florescu, I.: Bayesian fusion of thermal and visible spectra camera data for region based tracking with rapid background adaptation. In: 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 192–199. IEEE (2012)
16. Xiao, J., Stolkin, R., Oussalah, M., Leonardis, A.: Continuously adaptive data fusion and model relearning for particle filter tracking with multiple features. IEEE Sens. J. **16**(8), 2639–2649 (2016)
17. Possegger, H., Mauthner, T., Bischof, H.: In defense of color-based model-free tracking. In: CVPR, pp. 2113–2120 (2015)
18. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. PAMI **33**(9), 1806–1819 (2011)
19. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV, pp. 137–144. IEEE (2011)
20. Chen, S., Fern, A., Todorovic, S.: Multi-object tracking via constrained sequential labeling. In: CVPR, pp. 1130–1137. IEEE (2014)
21. Hare, S., Saffari, A., Torr, P.H.: Struck: Structured output tracking with kernels. In: ICCV, pp. 263–270. IEEE (2011)
22. Dehghan, A., Tian, Y., Torr, P.H., Shah, M.: Target identity-aware network flow for online multiple target tracking. In: CVPR, pp. 1146–1154 (2015)
23. Xiao, J., Stolkin, R., Leonardis, A.: Multi-target tracking in team-sports videos via multi-level context-conditioned latent behaviour models. In: BMVC (2014)
24. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003)
25. Myung, I.J.: Tutorial on maximum likelihood estimation. J. Math. Psychol. **47**(1), 90–100 (2003)
26. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33765-9_50
27. Mei, X., Ling, H.: Robust visual tracking using L1 minimization. In: ICCV, pp. 1436–1443. IEEE (2009)
28. Xiao, J., Oussalah, M.: Collaborative tracking for multiple objects in the presence of inter-occlusions. TCSVT **26**(2), 304–318 (2015)
29. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. PAMI **37**(3), 583–596 (2015)
30. Li, A., Lin, M., Wu, Y., Yang, M.H., Yan, S.: NUS-PRO: A new visual tracking challenge. PAMI **38**(2), 335–349 (2015)
31. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: CVPR, pp. 1838–1845. IEEE (2012)
32. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: ICCV, pp. 3074–3082(2015)