MTS: A Multiple Temporal Scale Tracker Handling Occlusion and Abrupt Motion Variation

Muhammad Haris Khan(✉), Michel F. Valstar, and Tony P. Pridmore

Computer Vision Laboratory, School of Computer Science,
University of Nottingham, Nottingham, UK
{psxmhk,michel.valstar,tony.pridmore}@nottingham.ac.uk

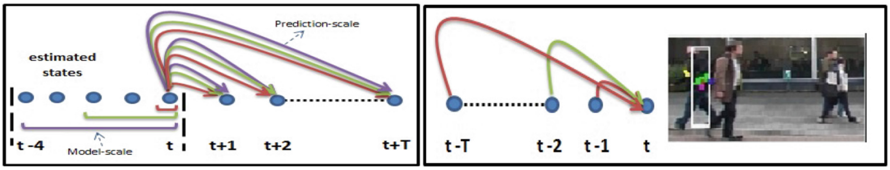
Abstract. We propose visual tracking over multiple temporal scales to handle occlusion and non-constant target motion. This is achieved by learning motion models from the target history at different temporal scales and applying those over multiple temporal scales in the future. These motion models are learned online in a computationally inexpensive manner. Reliable recovery of tracking after occlusions is achieved by extending the bootstrap particle filter to propagate particles at multiple temporal scales, possibly many frames ahead, guided by these motion models. In terms of the Bayesian tracking, the prior distribution at the current time-step is approximated by a mixture of the most probable modes of several previous posteriors propagated using their respective motion models. This improved and rich prior distribution, formed by the models learned and applied over multiple temporal scales, further makes the proposed method robust to complex target motion through covering relatively large search space with reduced sampling effort. Extensive experiments have been carried out on both publicly available benchmarks and new video sequences. Results reveal that the proposed method successfully handles occlusions and a variety of rapid changes in target motion.

1 Introduction

Visual tracking is one of the most important unsolved problems in computer vision. Though it has received much attention, no framework has yet emerged which can robustly track across a broad spectrum of real world settings. Two major challenges for trackers are abrupt variations in target motion and occlusions. In some applications, e.g. video surveillance and sports analysis, a target may undergo abrupt motion changes and be occluded at the same time.

While many solutions to the occlusion problem have been proposed, it remains unsolved. Some methods [1–3] propose an explicit occlusion detection and handling mechanism. Reliable detection of occlusion is difficult in practice, and often produces false alarms. Other methods, e.g. those based on adaptive appearance models [4,5], use statistical reasoning to handle occlusions indirectly, by learning how appearance changes over time. Occlusions can, however, contaminate the appearance models, as such methods use blind update strategies.

D. Cremers et al. (Eds.): ACCV 2014, Part V, LNCS 9007, pp. 476–492, 2015.
DOI: 10.1007/978-3-319-16814-2_31



(a) Multiple motion models are learned from the recent history of estimated states at different temporal scales, and each model is applied to multiple temporal scales in the future.

(b) This means that, when determining target state, multiple sets of motion models are available to make predictions. Each set includes models learnt at multiple model-scales. In the proposed framework one model per set is selected to propagate particles.

Fig. 1. Visual tracking over multiple temporal scales.

Abruptly varying motion can be addressed using a single motion model with a large process noise. This approach requires large numbers of particles and is sensitive to background distractors. Alternative approaches include efficient proposals [6], or hybrid techniques with hill climbing methods [7] to allocate particles close to the modes of the posterior. These approaches can, however, be computationally expensive.

We propose a new tracking method that is capable of implicitly coping with partial and full target occlusion and non-constant motion. To recover from occlusion we employ a flexible prediction method, which estimates target state at temporal scales similar to the expected maximum duration of likely occlusions. To achieve this, motion models are learnt at multiple model-scales and used to predict possible target states at multiple prediction-scales ahead in time. The model-scale is the duration of a sequence of recently estimated target states over which a motion model is learnt. The prediction-scale is the temporal distance, measured in frames of the input image sequence, over which a prediction is made. Reliable recovery of tracking after occlusions is achieved by extending the bootstrap particle filter to propagate particles to multiple prediction-scales, using models learnt at multiple model-scales. Figure 1 summarises the approach.

The proposed framework can handle variable motion well due to the following: In predictive tracking, learnt motion models describe the recent history of target state —the most recent section of the target's path across the image plane. Trackers using, for example, a single linear motion model effectively represent target path as a straight line. By building multiple motion models at multiple model scales, the proposed framework maintains a much richer description of target path. The diverse set of models produced captures at least some of the complexity of that path and, when used to make predictions, the model set represents variation in target motion better than any single model.

The contributions of this work are three-fold. **(1)** We propose and evaluate the idea of tracking over multiple temporal scales to implicitly handle occlusions of variable lengths and achieve robustness to non-constant target motion. This is accomplished by learning motion models at multiple model-scales and applying them over multiple prediction-scales. Consequently, the proposed framework does not require an explicit occlusion detection, which could be difficult to achieve reliably in practice. **(2)** We propose a simple but generic extension of the bootstrap particle filter to search around the predictions generated by the motion models. **(3)** Current trackers typically adopt a first-order Markov Chain assumption, and predict a target's state at time t using only its state at time $t-1$. That is, they all work on a single temporal scale i.e. $[t-1, t]$. We propagate important part of some recently estimated posteriors to approximate prior distribution at the current time-step through combining the above two proposals in a principled way. The resulting formulation is a tracker operating at multiple temporal scales that has not been proposed before to the best of our knowledge.

2 Related Work

Occlusion handling may be explicit or implicit. Implicit approaches can be divided into two categories. The first is based on adaptive appearance models which use statistical analysis [4,5,8] to reason about occlusion. The appearance models can, however, become corrupt during longer occlusions due to the lack of an intelligent update mechanism. Approaches in the second category divide the target into patches and either use a voting scheme [9] or robust fusion mechanism [10] to produce a tracking result. These can, however, fail when the number of occluded patches increases. The proposed approach also handles occlusion implicitly, but using a fixed and very simple appearance model.

Explicit occlusion handling requires robust occlusion detection. Collins et al. [1] presented a combination of local and global mode seeking techniques. Occlusion detection was achieved with a naive threshold based on the value of the objective function used in local mode seeking. Lerdsudwichai et al. [2] detected occlusions by using an occlusion grid with a drop in similarity value. This approach can produce false alarms because the required drop in similarity could occur due to natural appearance variation. To explicitly tackle occlusions, Kwak et al. [3] trained a classifier on the patterns of observation likelihoods in a completely offline manner. In [11,12], an occlusion map is generated by examining trivial coefficients, this is then used to determine the occlusion state of a target candidate. Both these methods are prone to false positives where it is hard to separate the intensity of the occluding object from small random noise. The proposed approach here does not detect occlusions explicitly, as it is difficult to achieve reliably.

Some approaches address domain-specific occlusion of known target types. Lim et al. [13] propose a human tracking system based on learning dynamic appearance and motion models. A three-dimensional geometric hand model was proposed by Sudderth et al. [14] to reason about occlusion in a non-parametric belief propagation tracking framework. Others [15,16] attempt to overcome occlusion using multiple cameras. As most videos are shot with a single camera, and multiple cameras bring additional costs; this is not a generally applicable solution. Furthermore, a domain-agnostic approach is more widely applicable.

Recently, some methods exploited context along with target description [17–19], and a few exploited detectors [20,21] to overcome occlusions. Context-based methods can tackle occlusions, but rely on the tracking of auxiliary objects.

Approaches based on detector could report false positives in the presence of distractors, causing the tracker to fail. Our approach does not search the whole image space, instead multiple motion models define relatively limited search spaces of variable size where there is high target probability. This results in reduced sampling effort and lower vulnerability to distractors.

When target motion is difficult to model, a common solution is to use a single motion model with a large process noise. Examples of such models are random-walk (RW) [7,22] and nearly constant velocity (NCV) [23,24]. Increased process noise demands larger numbers of particles to maintain accurate tracking, which increases computational expense.

One approach to the increased variance in estimation caused by high process noise is to make an efficient and informed proposal distribution. Okuma et al. [6] designed a proposal distribution that mixed hypotheses generated by an AdaBoost detector and a standard autoregressive motion model to guide a particle filter based tracker. Reference [25] formulated a two-stage dynamic model to improve the accuracy and efficiency of the bootstrap PF, but their method fails during frequent spells of non-constant motion. Kwon and Lee [8] sampled motion models generated from the recent sampling history to enhance the accuracy and efficiency of MCMC based sampling process. We also learn multiple motion models, but at different model-scales instead of a single scale and use recently estimated states history in comparison to sampling history.

Several attempts have been made to learn motion models offline. Isard and Blake [26] use a hardcoded finite state machine (FSM) to manage transitions between a small set of learned models. Madrigal et al. [27] guide a particle filter based target tracker with a motion model learned offline. Pavlovic et al. [28] switch between motion models learned from motion capture data. Their approach is application specific, in that it learns only human motion. Reference [29] classifies videos into categories of camera motion and predicts the right specialist motion model for each video to improve tracking accuracy, while we learn motion models over multiple temporal scales in an online manner to generate better predictions. An obvious limitation of offline learning is that models can only be used to track the specific class of targets for which they are trained.

To capture abrupt target motion, which is difficult for any motion model, [30] combined an efficient sampling method with an annealing procedure, [31] selects easy-to-track frames first and propagates density from all the tracked frames to a new frame through a patch matching technique, and [32] introduced a new sampling method into the Bayesian tracking. Our proposed method tries to capture reasonable variation in the target's path.

Two approaches that at first glance appear similar to ours are [33,34]. Mikami et al. [33] use the entire history of estimated states to generate a prior distribution over the target state at immediate and some future time-steps, though the accuracy of these prior distributions relies on strict assumptions. In [34], offline training is required prior to tracking and thus it cannot be readily applied to track any object. Our approach learns multiple simple motion models at relatively short temporal scales in a completely online setting, and each model predicts the target state at multiple temporal scales in the future.

In contrast to previous work, we learn motion models over multiple model-scales, whose predictions are pooled over multiple prediction-scales to define the search space of a single particle filter. Hence, this is an online learning approach not restricted to any specific target class, and a novel selection criterion selects suitable motion models without the need for a hardcoded FSM.

3 Bayesian Tracking Formulation

Our aim is to find the best state of the target at time t given observations up to t. State at time t is given by $\mathbf{X}_t = \{X_t^x, X_t^y, X_t^s\}$,where X_t^x, X_t^y, and X_t^s represent the x, y location and scale of the target, respectively. In a Bayesian formulation, our solution to tracking problem comprises two steps: update(1), and prediction (2).

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \propto p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}). \tag{1}$$

where $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$ is the posterior probability given the state $\mathbf{X}_t$ at time t, and observations $\mathbf{Y}_{1:t}$ up to t. $p(\mathbf{Y}_t|\mathbf{X}_t)$ denotes the observation model.

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}) = \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})d\mathbf{X}_{t-1}. \tag{2}$$

where $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1})$ is the prior distribution at time t, and $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is a motion model.

In this work, we improve the accuracy of the posterior distribution at a given time t by improving the prior distribution. Here, the prior distribution is approximated by a mixture of the most probable modes of T previous posteriors propagated by the T selected motion models, which are generated using information from up to T frames ago. The Eq. 2 in the standard Bayesian formulation can now be written as:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}) = \int_{k=1}^{k=T} \int_{\mathbf{X}_{t-k}} p(\mathbf{X}_t|\mathbf{X}_{t-k})p(\tilde{\mathbf{X}}_{t-k})d\mathbf{X}_{t-k}dk. \tag{3}$$

where $p(\mathbf{X}_t|\mathbf{X}_{t-k})$ is the motion model selected at time t from a set of motion models learned at time $t-k$, and $p(\tilde{\mathbf{X}}_{t-k}) \subset p(\mathbf{X}_{t-k}|\mathbf{Y}_{1:t-k})$ is the most probable mode (approximated via particles) of the posterior at time $t-k$. A relatively rich and improved prior distribution in Eq. 3 allows handling occlusions and abrupt motion variation in a simple manner without resorting to complex appearance models and exhaustive search methods.

The best state of the target $\hat{\mathbf{X}}_t$ is obtained using Maximum a Posteriori (MAP) estimate over the N_t weighted particles which approximate $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$,

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t^{(i)}} p(\mathbf{X}_t^{(i)}|\mathbf{Y}_{1:t})\, for\, i = 1, ..., N_t, \tag{4}$$

where $\mathbf{X}_t^{(i)}$ is the i_{th} particle.

4 Proposed Method

4.1 A Multiple Temporal Scale Framework

To reliably recover the target after occlusion and achieve robustness to non-constant motion, we introduce the concept of learning motion models at a range of model-scales, and applying those over multiple prediction-scales. Furthermore, we contribute a simple but powerful extension of the bootstrap particle filter to search around the predictions generated by the motion models.

The core idea is to construct an improved and rich prior distribution at each time-point by combining sufficient particle sets that at least one set will be valid and allow recovery from occlusion and robustness to non-constant motion. A valid particle set is the most probable mode of an accurate estimation of the posterior probability from some previous time-point, propagated by a motion model generated over an appropriate model-scale and unaffected by occlusion.

Learning and Predicting Motion Over Multiple Temporal Scales. Simple motion models are learned over multiple model-scales and are used to make state predictions over multiple prediction-scales. A simple motion model is characterized by a polynomial function of order d, and represented by $\mathbf{M}$. $\mathbf{M}$ is learned at a given model-scale separately for the x-location, y-location, and scale s of the target's state.[1] This learning also considers how well each state is estimated in a given sequence and how far it is from the most recently estimated state [25]. For instance, an $\mathbf{M}$ of order 1, learned at model-scale m, predicts a target's x-location at time t as:

$$\tilde{x}_t = \beta_o^m + \beta_1^m t, \tag{5}$$

where β_1 is the slope, and β_o the intercept. Model parameters can be learned inexpensively via weighted least squares.

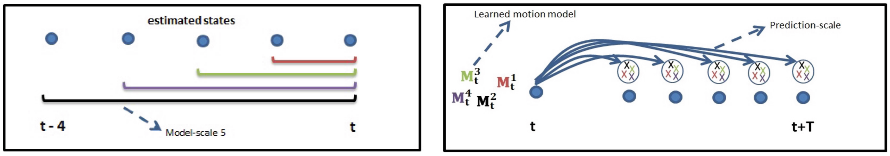
A set of learned motion models at time t is represented by $\mathbf{M}_t^{j=1,\ldots,|\mathbf{M}_t|}$, where $|.|$ is the cardinality of the set. Each model predicts target state $l(\tilde{x}, \tilde{y}, \tilde{s})$ at T prediction-scales. See Figs. 2a and b for an illustration of learning and prediction.

Model Set Reduction. The aim of model set reduction is to establish search regions for the particle filter in which there is a high probability of target being present. This in turn will reduce the sampling effort as search regions corresponding to all the predictions no longer need to be searched.

Suppose there are T sets of motion models available at time t, one from each of T previous time-steps. Each set of models at time t is represented by its corresponding set of predictions. The most suitable motion model from each set is selected as follows.

Let us denote $G = |\mathbf{M}_t|$, and let $\mathbf{l}_t^k = \{l_t^{j,k} | j = 1, ..., G\}$ represent a set of states predicted by G motion models learned at time $t - k$, where $l_t^{j,k}$ denotes

[1] To demonstrate the basic idea of the proposed approach and for the sake of simplicity, x, y, and s part of the target state are considered uncorrelated. They may be correlated, and taking this into account while learning might produce improved models. We would pursue this avenue in future work.



(a) During learning, multiple motion models are constituted at multiple model-scales using the recent history of estimated states at time t. In this figure, four linear motion models are learned over four different model-scales at time t. The four model-scales are 2,3,4, and 5.

(b) During prediction, a set of learned motion models are used to predict possible target states at T prediction-scales. In this figure, a set comprising four learned motion models is shown at time t. Each motion model predicts possible target state at T prediction-scales.

Fig. 2. Graphical illustration of what happens during learning and prediction.

the predicted state by j_{th} motion model learned at k_{th} previous time-step. As $k = 1, ..., T$, there are T sets of predicted states at time t (Fig. 3(a)). Now the most suitable motion model $\mathbf{R}_t^k$ is selected from each set using the following criterion:

$$\hat{l}_t^k = \arg\max_{l_t^{j,k}} p(\mathbf{Y}_t | l_t^{j,k}) \tag{6}$$

where $\hat{l}_t^k$ is the most suitable state prediction from the set $\mathbf{l}_t^k$, and $p(\mathbf{Y}_t | l_t^{j,k})$ measures the visual likelihood at the predicted state $l_t^{j,k}$. In other words, $\hat{l}_t^k$ is the most suitable state prediction of the most suitable motion model $\mathbf{R}_t^k$. For example, Fig. 3(b) shows the predicted state $\hat{l}_t^1$ of the most suitable motion model $\mathbf{R}_t^1$ chosen from 4 motion models learned at time $t-1$. After this selection process, the T sets of motion models are reduced to T individual models.

Propagation of Particles. In the bootstrap particle filter [35], the posterior probability at time $t-1$ is estimated by a set of particles $\mathbf{X}_{t-1}^{(i)}$ and their weights $\omega_{t-1}^{(i)}, \{\mathbf{X}_{t-1}^{(i)}, \omega_{t-1}^{(i)}\}_{i=1}^N$, such that all the weights in the particle set sum to one. The particles are resampled to form an unweighted representation of the posterior $\{\mathbf{X}_{t-1}^{(i)}, 1/N\}_{i=1}^N$. At time t, they are propagated using the motion model $p(\mathbf{X}_t | \mathbf{X}_{t-1})$ to approximate a prior distribution $p(\mathbf{X}_t | \mathbf{Y}_{t-1})$. Finally, they are weighted according to the observation model $p(\mathbf{Y}_t | \mathbf{X}_t)$, approximating the posterior probability at time t.

Here, particle sets not just from one previous time-step $(t-1)$, but from T previous time-steps are propagated to time t using the T selected motion models. When using first-order polynomial (linear) motion models the most suitable motion model $\mathbf{R}_t^k$ selected from those learnt at the k_{th} previous time-step will propagate a particle set from the k_{th} previous time-step as follows

$$X_{t,k}^x = X_{t-k}^x + g(\mathbf{R}_t^k)k + \mathcal{N}(0, \sigma_x^2 k), \tag{7}$$

where X^x is the horizontal part of the target state, $g()$ indicates the slope of the model, and $\mathcal{N}(0, \sigma_x)$ is a Gaussian distribution with zero-mean and σ_x^2 variance. For instance, in Fig. 3(c), the most suitable motion model $\mathbf{R}_t^1$, is used to propagate a particle set from time $t-1$ to time t.

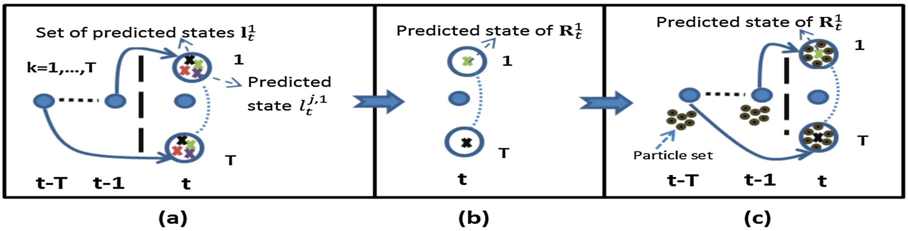

Set of predicted states $\mathbf{l}_t^1$
k=1,...,T
1
Predicted
state $l_t^{j,1}$
T
t-T
t-1
t
(a)
Predicted state of $\mathbf{R}_t^1$
1
T
t
(b)
Predicted state of $\mathbf{R}_t^1$
1
Particle set
T
t-T
t-1
t
(c)

Fig. 3. (a) Before model set reduction, there exist T different sets of predicted states at time t, where each set $\mathbf{l}_t^k$ comprises G states predicted by G motion models learned at k_{th} previous time-step. In Fig. 3(a), $\mathbf{l}_t^1$ is a set composed of 4 states predicted by 4 motion models learned at time $t-1$. (b) **Model Set Reduction.** T sets of motion models available at time t, represented by the corresponding T sets of predicted states, are reduced to T individual models. Figure 3(b) shows the predicted state $\hat{l}_t^1$ of the most suitable motion model $\mathbf{R}_t^1$ selected from 4 motion models learned at time $t-1$. (c) **Propagation of Particles.** T selected motion models, one from each of the T preceding time-steps, are used to propagate particle sets from T preceding time-steps to time t. In Fig. 3(c), the most suitable motion model $\mathbf{R}_t^1$, is used to propagate particle set from time $t-1$ to time t.

Propagation from the last T time-steps, generates T particle sets at time t. All particles are weighted using the observation model $p(\mathbf{Y}_t|\mathbf{X}_t)$ to approximate the posterior probability $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$. If the target was occluded for less than or equal to $T-1$ frames, it may be recovered by a set of particles unaffected by the occlusion. To focus on particles with large weights, and reduce computational cost, we retain the first N particles after the resampling step. The proposed framework is summarised in Algorithm 1.

5 Experimental Results

In the proposed method, the appearance model used in all experiments was the colour histogram used in [36].[2] The Bhattacharyya coefficient was used as the distance measure. As simple motion model the first-order polynomial (linear) model with model-scales of 2, 3, 4, and 5 frames was used (four models in total).

MTS-L denotes the proposed method applied over a first-order polynomial (linear) motion model (Algorithm 1). We also apply our proposed framework to the two-stage model of [25], which is denoted by MTS-TS, to show its generality[3].

[2] We investigate the power of using multiple temporal scales of motion model generation and application to deal with visual tracking problems related to occlusion and abrupt motion variation. To evaluate this hypothesis independently of the appearance model, a simple appearance model is used on purpose.

[3] MTS-TS is identical to MTS-L except that the propagation of particles takes place through a different model instead of the model proposed in Eq. 7 and the variance of the best state (estimated through particles) is reduced by combining it with the highest likelihood motion prediction. See the supplementary material for the details of this application.

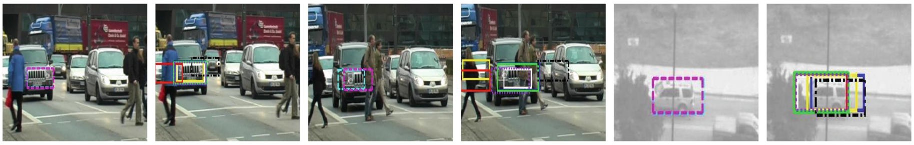
Algorithm 1. A Multiple Temporal Scale Tracker

```
Input: Let $\mathbf{W} = \{W_{t-1}, ..., W_{t-T}\}$ represent the resampled sets of particles after
estimation of the posterior from $T$ previous time-steps, where $W_{t-1} = \{\mathbf{X}_{t-1}^{(i)}, \frac{1}{N}\}_{i=1}^{N}$.
Output: Best state $\hat{\mathbf{X}}_t$ at time $t$.
  for  $k = 1$ to $T$
    for  $j = 1$ to $G$
      - Measure visual likelihood $p(\mathbf{Y}_t|l_t^{j,k})$, where $l_t^{j,k}$ denotes the predicted state
        at time $t$ by $j_{th}$ motion model from $k_{th}$ previous time-step.
    end
    - Select the most suitable motion model $\mathbf{R}_t^k$ at time $t$ using Eq. 6.
    - Propagate the particle set from $k_{th}$ previous time-step $W_{t-k} = \{\mathbf{X}_{t-k}^{(i)}, \frac{1}{N}\}_{i=1}^{N}$
      using Eq. 7 by taking the slope of selected motion model $\mathbf{R}_t^k$ to time $t$.
  end
- Assign weights to all the particles to approximate the posterior $\{\mathbf{X}_t^{(i)}, \omega_t^{(i)}\}_{i=1}^{N \times T}$.
- Calculate the best state $\hat{\mathbf{X}}_t$ using Eq. 4.
- Retain first N particles after the resampling step.
- Learn simple motion models using the recent history of estimated states.
```

In MTS-TS, the β parameter of the two-stage model was fixed at 10, giving high weight to the rigid velocity $\hat{v}$, estimated by the simple motion model, and very low weight to the internal velocity v. As a result, it becomes strongly biased towards the predicted location, but still allows some deviation.

We compared the proposed method to three baseline and seven state-of-the-art trackers. The first two baseline trackers, T_{RW} and T_{NCV}, were colour based particle filters from [36], but use different motion models. T_{RW} used a random-walk model while T_{NCV} used a nearly constant velocity model. The third baseline tracker T_{TS} was the two-stage dynamic model proposed by [25]. The parameters, K and β, in [25] were set to 5 and 10, respectively. The state-of-the-art trackers are SCM [37], ASLA [38], L1-APG [12], VTD [5], FragT [9], SemiBoost [21], and WLMCMC [30]. The minimum and maximum number of samples used for WLMCMC, VTD, SCM, ASLA, and L1-APG was 600 and 640, respectively. Our proposed tracker is implemented in MATLAB and runs at about 3 frames/sec with 640 particles. The source code and datasets (along with ground truth annotations) will be made available on the authors' webpages.

We chose state-of-the-art trackers keeping in view two important properties: their performance according to the CVPR'13 benchmark [39], and their ability to handle occlusions (partial and full) and abrupt motion variations. SCM and ASLA both have top ranked performance on the CVPR'13 benchmark. SCM combines a sparsity based classifier with a sparsity based generative model and has a occlusion handling mechanism, while ASLA is based on a local sparse appearance model and is robust to partial occlusions. In L1-APG, the coupling of L1 norm minimization and an explicit occlusion detection mechanism makes it robust to partial as well as full occlusions. The integration of two motion models having different variances with a mixture of template-based object models lets VTD explore a relatively large search space, while remaining robust to a wide



(a) TU-Cr♯46 (b) TU-Cr♯46 (c) TU-Cr♯124 (d) TU-Cr♯124 (e) car♯169 (f) car♯169

Fig. 4. Tracking through multiple partial occlusions. MTS-TS(magenta), MTS-L (cyan), SCM(green) FragT(white), Semi(yellow), L1-APG(blue), VTD(red), WLMCMC (black), and ASLA(purple) (Color figure online).

range of appearance variations. FragT was chosen because its rich, patch-based representation makes it robust to partial occlusion. SemiBoost was picked as it searches the whole image space once its tracker loses target, and thus, it can re-locate the target after full occlusions. WLMCMC searches the whole image space by combining an efficient sampling strategy with an annealing procedure that allows it to capture abrupt motion variations quite accurately and re-locate the target after full occlusions.

Eleven video sequences were used. Seven are publicly available (*PETS 2001 Dataset 1*[4], *TUD-Campus* [40], *TUD-Crossing* [40], *Person* [41], *car*[5] [39], *jogging* [39], and *PETS 2009 Dataset S2*[6]) and four are our own (*squash*, *ball1*,*ball2*, and *toy1*). All involve frequent short and long term occlusions (partial and full) and/or variations in target motion. We used three metrics for evaluation: centre location error, Pascal score [42], and precision at a fixed threshold of 20 pixels [43].

5.1 Comparison with Competing Methods

Quantitative Evaluation: Table 1 summarises tracking accuracy obtained from sequences in which the target is occluded. MTS-L outperformed competing methods in most sequences, because it efficiently allocated particles to overcome occlusions. VTD performed badly because inappropriate appearance model updates during longer occlusions causes drift from which it cannot recover. Although SemiBoost uses explicit re-detection once the target is lost, its accuracy was low due to false positive detections. With the ability to search the whole image space using an efficient sampling scheme, WLMCMC produced the lowest error in the *TUD-Campus* and *jogging* sequences. In sequences containing partial occlusions (Fig. 4), SCM produced the lowest error in the *car* sequence, while both SCM and L1-APG had the best performance in the *TUD-Crossing* sequence. SCM uses a sparse based generative model that considers spatial relationship among local patches with an occlusion handling scheme, and L1-APG

[4] *PETS 2001 Dataset 1* is available from http://ftp.pets.rdg.ac.uk/.

[5] We downsampled original *car* sequence by a factor of 3 to have partially low frame rate.

[6] *PETS 2009 Dataset S2* is available from http://www.cvg.rdg.ac.uk/PETS2009/.

employs a robust minimization model influenced by an explicit occlusion detection mechanism. Thus, both these approaches are quite effective in overcoming partial occlusions. In contrast, MTS-L and MTS-TS use a very simple, generic appearance model, and no explicit occlusion handling mechanism[7].

Table 1. Tracking accuracy in the presence of occlusion

(a) Mean centre location error in pixels is given, averaged over all frames of all videos showing occlusions. Each tracker was run five times and the results were averaged. The best results are marked in bold. T denotes the prediction-scales, and N is the number of particles propagated from $t-k$ to t in our proposed method. N is fixed at 20, and N_t is the total number of particles accumulated at time t in our proposed method. The number of particles used in baseline trackers was equal to N_t.

Sequence	SCM	ASLA	L1	VTD	Semi	FragT	WL	T_{NCV}	T_{RW}	T_{TS}	MTS-L	MTS-TS	T	N_t=N×T
ball2	76	71	71	66	78	106	37	91	71	125	17	**16**	32	640
TUD-Camp	186	180	100	186	61	112	**22**	141	119	31	24	**22**	9	180
TUD-Cross	**2**	6	**2**	63	62	5	50	43	75	106	25	21	25	500
PETS 2001	61	63	60	83	114	67	90	43	131	112	25	**21**	32	640
Person	91	80	103	85	177	84	25	90	33	95	10	**8**	20	400
PETS 2009	35	13	81	94	29	10	91	75	37	56	7	**6**	14	280
car	**8**	31	31	47	38	15	28	37	43	87	25	25	20	400
toy1	88	85	111	98	99	107	30	74	134	76	**21**	22	30	600
jogging	110	104	45	70	30	94	**19**	27	24	100	25	24	20	400

(b) A(B): A - the percentage of correctly tracked frames based on Pascal Score [42]; B - Precision at a fixed threshold of 20 pixels. Pascal score is computed by assessing to what extent the tracking template overlaps the ground truth template as a ratio. If the Pascal score is greater than 0.5 in a certain frame, that frame is counted as a correctly tracked frame. Precision is computed by dividing the number of frames, where estimated target location was not beyond the fixed threshold distance of 20 pixels of the ground truth, by the total number of frames in a video sequence. The best results are marked in bold.

Sequence	SCM	ASLA	L1	VTD	Semi	FragT	WL	MTS-L	MTS-TS
ball2	12(0.21)	9(0.17)	7(0.21)	9(0.11)	7(0.13)	9(0.09)	28(0.53)	31(0.8)	**36**(**0.8**)
TUD-Camp	14(0.17)	10(0.14)	19(0.21)	25(0.25)	38(0.34)	27(0.27)	46(**0.61**)	55(0.57)	**57**(0.46)
TUD-Cross	**100**(**1**)	99(0.9)	**100**(**1**)	24(0.23)	41(0.42)	87(1)	25(0.23)	61(0.59)	69(0.65)
PETS 2001	23(0.33)	23(0.27)	22(0.25)	20(0.25)	17(0.2)	16(0.31)	19(0.52)	58(0.65)	**66**(**0.7**)
Person	45(0.46)	44(0.45)	10(0.12)	43(0.45)	20(0.2)	38(0.41)	49(0.86)	79(0.93)	**80**(**0.94**)
PETS 2009	26(0.36)	36(0.7)	21(0.26)	21(0.21)	27(0.45)	65(0.73)	7(0.23)	70(0.97)	**71**(**0.96**)
car	**92**(**0.93**)	62(0.64)	66(0.65)	66(0.65)	55(0.46)	80(0.76)	62(0.52)	71(0.72)	73(0.72)
toy1	18(0.19)	19(0.2)	15(0.15)	16(0.18)	16(0.18)	3(0.09)	**49**(**0.8**)	43(0.8)	38(0.78)
jogging	21(0.22)	22(0.22)	21(0.21)	22(0.22)	**60**(**0.61**)	21(0.21)	42(**0.61**)	20(0.44)	21(0.45)

Tracking accuracy was also measured when the target was occluded and underwent motion variation at the same time (Tables 2a and b). MTS-L produced higher accuracy than the other methods. The allocation of particle sets with different spreads from multiple prediction scales in regions having probable local maxima lets MTS-L capture increased search space with relatively smaller sampling effort. VTD performed well in *squash* sequence because it combines two motion models of different variances to form multiple basic trackers which search a large state space efficiently. WLMCMC produced the second best accuracy on *ball1* as it searches the whole image space using an efficient sampling mechanism to capture abrupt target motion.

[7] We admit that a more complex system complete with more advanced appearance models would obtain a higher overall tracking accuracy, but we believe that for the sake of scientific evidence finding employing such a system would obfuscate attribution of our experimental results to the original hypothesis.

Table 2. Accuracy through simultaneous motion variation and occlusion

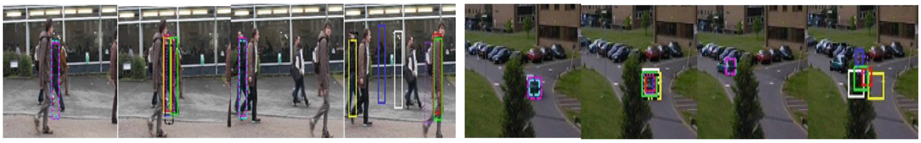
(a) Mean centre location error (pixels).

Sequence	SCM	ASLA	L1	VTD	Semi	FragT	WL	T_{NCV}	T_{RW}	T_{TS}	MTS-L	MTS-TS	T	N_t=N×T
squash	40	34	60	20	68	35	22	27	52	41	12	**10**	5	100
ball1	91	96	124	69	66	188	23	74	87	98	15	**14**	14	280

(b) A(B): A - the percentage of correctly tracked frames based on Pascal Score; B - Precision at a fixed threshold of 20 pixels.

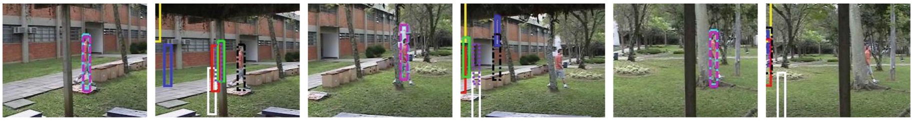
Sequence	SCM	ASLA	L1	VTD	Semi	FragT	WL	MTS-L	MTS-TS
squash	60(0.62)	38(0.56)	9(0.11)	68(0.78)	44(0.7)	37(0.5)	50(0.75)	71(0.92)	**75**(**0.96**)
ball1	6(0.06)	3(0.04)	2(0.05)	19(0.22)	19(0.33)	2(0.02)	35(0.79)	40(0.83)	**41**(**0.89**)

Qualitative Evaluation: Tracking is particularly difficult when the time between consecutive occlusions is small. In *TUD-Campus*, the tracked person suffers two occlusions only 17 frames apart (Fig. 5a). MTS-L and WLMCMC recover the target after each occlusion, while other methods fail due to incorrect appearance model updates, or being distracted by the surrounding clutter. Video surveillance data often requires tracking through partial and/or full occlusions. In the *PETS 2001 Dataset 1* sequence (Fig. 5b) the target (car) first stays partially occluded for a considerable time, and is then completely occluded by a tree. MTS-L successfully re-acquires the target. Occlusions of varying lengths are common in real-world tracking scenarios. In the *person* sequence, a person moves behind several trees and is shot with a moving camera. As shown in Fig. 6, competing methods lose the target after first occlusion(Frame # 238) or second occlusion(Frame # 329), while MTS-L shows robustness in coping with varying lengths of occlusions.



(a) TUD-Campus♯12♯39 (b) PETS'01♯38♯92

Fig. 5. Tracking results in a crowded (a) and a surveillance environment (b).



(a) Person♯238 (b) Person♯238 (c) Person♯329 (d) Person♯329 (e) Person♯457 (f) Person♯457

Fig. 6. Tracking results with occlusions of different lengths.

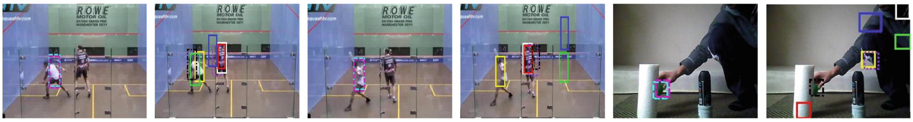

(a) Squash♯74 (b) Squash♯74 (c) Squash♯274 (d) Squash♯274 (e) ball1♯792 (f) ball1♯792

Fig. 7. Tracking results in case of motion variations and frequent occlusions.

The ability of MTS-L to cope with simultaneous occlusion and non-constant target motion was tested by making two challenging sequences: *squash* and *ball1*. In these sequences, the target accelerates, decelerates, changes direction suddenly, and is completely occluded multiple times. Figure 7 illustrates tracking results. MTS-L provided more accurate tracking than the other methods on both sequences. This is because the efficient allocation of particles at multiple prediction-scales allows a wider range of target motion to be handled. WLM-CMC shows good accuracy in the *ball1* sequence as it is aimed at handling abrupt target motion.

5.2 Analysis of the Proposed Framework

Without Multiple Prediction-Scales. The proposed framework was tested without employing multiple prediction-scales. We designed MTSWPS-L in which the target state is predicted only 1 frame ahead i.e. $T = 1$. For evaluation, at first, the number of particles in MTSWPS-L was kept equal to N_t and the process noise σ_{xy} was same as used for MTS-L between two consecutive time-steps. To analyze further, later, both the number of particles N_t and the process noise σ_{xy} were doubled and tripled. Figure 8(left) reveals the performance of the proposed framework with and without multiple prediction-scales in five video sequences

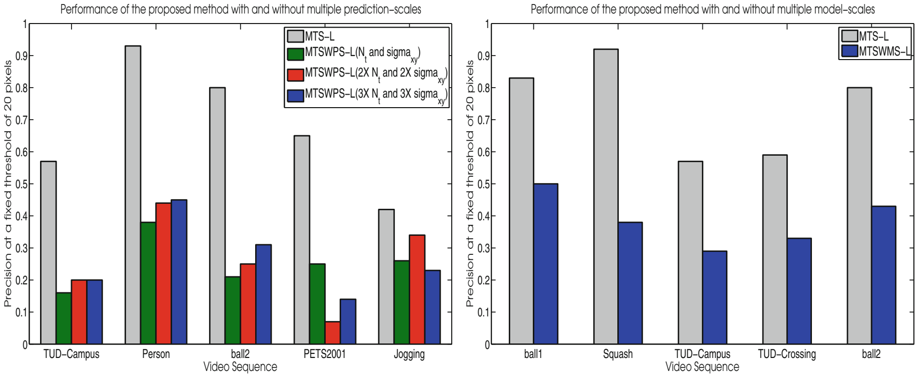


Fig. 8. (left)Performance of the proposed framework with and without multiple prediction-scales. (right)Performance of the proposed framework with and without multiple model-scales.

involving occlusions. As can be seen, MTSWPS-L has poor performance compared to MTS-L in all 5 sequences even after increasing the sampling effort and the process noise by three times of the original. Therefore, we can say that operation over multiple prediction-scales allows the proposed method to reliably handle occlusions in a principled way.

Without Multiple Model-Scales. The proposed framework was also analyzed without learning over multiple model-scales. MTSWMS-L denotes the proposed framework in which a linear motion model is learned over model-scale 2 only. As a result, there is no need to select models from each of the previous time-steps at the current time-step since only 1 model is learned over a single model-scale. As can be seen in Fig. 8(right), MTS-L has superior performance over MTSWMS-L in all 5 sequences. This shows that by constructing motion models over multiple model-scales MTS-L maintains a richer description of the target's path, which is not possible with a single scale model. Furthermore, this diverse set of models produces temporal priors that ultimately develops into a rich prior distribution required for reliably recovery of tracking after occlusions.

Experimental results show the robust performance of the proposed framework during occlusions, but it can fail when faced with very long duration occlusions. In addition, it can be distracted by visually similar objects after occlusion, if the state estimations during the period of occlusion are poor.

6 Conclusion

We propose a tracking framework that combines motion models learned over multiple model-scales and applied over multiple prediction-scales to handle occlusion and variation in target motion. The core idea is to combine sufficient particle sets at each time-point that at least one set will be valid, and allow recovery from occlusion and/or motion variation. These particle sets are not, however, simply spread widely across the image: each represents an estimation of the posterior probability from some previous time-point, predicted by a motion model generated over an appropriate model-scale.

The proposed method has shown superior performance over competing trackers in challenging tracking environments. That there is little difference between results obtained using linear and two-stage motion models suggests that this high level of performance is due to the framework, rather than its components.

References

1. Yin, Z., Collins, R.T.: Object tracking and detection after occlusion via numerical hybrid local and global mode-seeking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
2. Lerdsudwichai, C., Abdel-Mottaleb, M., Ansari, A.: Tracking multiple people with recovery from partial and total occlusion. Pattern Recogn. **38**, 1059–1070 (2005)
3. Kwak, S., Nam, W., Han, B., Han, J.H.: Learning occlusion with likelihoods for visual tracking. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1551–1558. IEEE (2011)

4. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. **77**, 125–141 (2008)
5. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1269–1276. IEEE (2010)
6. Okuma, K., Taleghani, A., de Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: multitarget detection and tracking. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)
7. Naeem, A., Pridmore, T.P., Mills, S.: Managing particle spread via hybrid particle filter/kernel mean shift tracking. In: BMVC, pp. 1–10 (2007)
8. Kwon, J., Lee, K.M.: Tracking by sampling trackers. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1195–1202. IEEE (2011)
9. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 798–805. IEEE (2006)
10. Han, B., Davis, L.S.: Probabilistic fusion-based parameter estimation for visual tracking. Comput. Vis. Image Underst. **113**, 435–445 (2009)
11. Mei, X., Ling, H., Wu, Y., Blasch, E., Bai, L.: Minimum error bounded efficient l1 tracker with occlusion detection. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1257–1264. IEEE (2011)
12. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust l1 tracker using accelerated proximal gradient approach. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1830–1837. IEEE (2012)
13. Lim, H., Camps, O.I., Sznaier, M., Morariu, V.I.: Dynamic appearance modeling for human tracking. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 751–757. IEEE (2006)
14. Sudderth, E.B., Mandel, M.I., Freeman, W.T., Willsky, A.S.: Distributed occlusion reasoning for tracking with nonparametric belief propagation. In: Advances in Neural Information Processing Systems, pp. 1369–1376 (2004)
15. Dockstader, S.L., Tekalp, A.M.: Multiple camera tracking of interacting and occluded human motion. Proc. IEEE **89**, 1441–1455 (2001)
16. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 267–282 (2008)
17. Grabner, H., Matas, J., Van Gool, L., Cattin, P.: Tracking the invisible: Learning where the object might be. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1285–1292. IEEE (2010)
18. Yang, M., Wu, Y., Hua, G.: Context-aware visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1195–1209 (2009)
19. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1177–1184. IEEE (2011)
20. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1409–1422 (2012)
21. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
22. Perez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proc. IEEE **92**, 495–513 (2004)
23. Shan, C., Tan, T., Wei, Y.: Real-time hand tracking using a mean shift embedded particle filter. Pattern Recogn. **40**, 1958–1970 (2007)

24. Pernkopf, F.: Tracking of multiple targets using online learning for reference model adaptation. IEEE Trans. Syst. Man Cybern., B **38**, 1465–1475 (2008)
25. Kristan, M., Kovačič, S., Leonardis, A., Perš, J.: A two-stage dynamic model for visual tracking. IEEE Trans. Syst. Man Cybern. B **40**, 1505–1520 (2010)
26. Isard, M., Blake, A.: A mixed-state condensation tracker with automatic model-switching. In: Sixth International Conference on Computer Vision, pp. 107–112. IEEE (1998)
27. Madrigal, F., Rivera, M., Hayet, J.-B.: Learning and regularizing motion models for enhancing particle filter-based target tracking. In: Ho, Y.-S. (ed.) PSIVT 2011, Part II. LNCS, vol. 7088, pp. 287–298. Springer, Heidelberg (2011)
28. Pavlovic, V., Rehg, J.M., MacCormick, J.: Learning switching linear models of human motion. In: NIPS, Citeseer, pp. 981–987 (2000)
29. Cifuentes, C.G., Sturzel, M., Jurie, F., Brostow, G.J., et al.: Motion models that only work sometimes. In: British Machive Vision Conference (2012)
30. Kwon, J., Lee, K.M.: Tracking of abrupt motion using wang-landau monte carlo estimation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 387–400. Springer, Heidelberg (2008)
31. Hong, S., Kwak, S., Han, B.: Orderless tracking through model-averaged posterior estimation. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2296–2303. IEEE (2013)
32. Zhou, X., Lu, Y., Lu, J., Zhou, J.: Abrupt motion tracking via intensively adaptive markov-chain monte carlo sampling. IEEE Trans. Image Process. **21**, 789–801 (2012)
33. Mikami, D., Otsuka, K., Yamato, J.: Memory-based particle filter for face pose tracking robust under complex dynamics. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 999–1006. IEEE (2009)
34. Li, Y., Ai, H., Lao, S., et al.: Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
35. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans. Signal. Proc. **50**, 174–188 (2002)
36. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
37. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1838–1845. IEEE (2012)
38. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1822–1829. IEEE (2012)
39. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418. IEEE (2013)
40. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
41. Dihl, L., Jung, C.R., Bins, J.: Robust adaptive patch-based object tracking using weighted vector median filters. In: 2011 24th SIBGRAPI Conference on Graphics, Patterns and Images (Sibgrapi), pp. 149–156. IEEE (2011)

42. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: Prost: Parallel robust online simple tracking. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 723–730. IEEE (2010)
43. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 983–990. IEEE (2009)