Sun *et al. EURASIP Journal on Advances in Signal Processing* (2020) 2020:9
https://doi.org/10.1186/s13634-020-0665-x

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Object contour tracking via adaptive data-driven kernel

Xin Sun[1], Wei Wang[1], Dong Li[2*], Bin Zou[3] and Hongxun Yao[3]

Abstract

We present a novel approach to non-rigid object tracking in this paper by deriving an adaptive data-driven kernel. In contrast with conventional kernel-based trackers which suffer from the constancy of kernel shape as well as scale and orientation selection problem when the tracking targets are changing in size, the adaptive kernel can robustly achieve the adaptation to target variation and act toward the actual target contour simultaneously with the mean shift iterations. Level set technique is novelly introduced to the mean shift sample space to both cope with insufficient low-level information and implement the adaptive kernel evolution and update. Since the active contour model is designed to drive the kernel constantly to the direction that maximizes the appearance similarity, this adaptive kernel can continually seize the target shape to give a better estimation bias and produce accurate shift of the mean. Finally, accurate target region can successfully avoid the performance loss stemmed from pollution of background pixels hiding inside the kernel and qualify the samples fed the next time step. Experimental results on a number of challenging sequences validate the effectiveness of the technique.

1 Introduction

Object tracking is a challenging research topic in the field of computer vision. In previous literature, numerous approaches have been dedicated to compute the translation of an object in consecutive frames [1–4], among which the mean shift methods show impressive performances and have received a considerable amount of attention. As a nonparametric density estimator firstly appeared in [5], mean shift iteratively computes the nearest mode of a point sample distribution. Then, it was applied by Comaniciu [6] to object tracking where the cost function between two color histograms is minimized through the mean shift iterations.

Despite its promising performance [7–10], there is a significant problem facing the traditional mean shift, i.e., the unclear kernel scale selection mechanism. Since the scale of mean shift kernel directly determines the size of the window within which sample weights are examined and affect the amount of kernel shift, it is a crucial parameter for the mean shift algorithm. However, there is currently no sound mechanism for choosing this scale maturely. The intuitive approach is to search for the best scale by testing different kernel bandwidths and selecting the one maximizing appearance similarity. This kind of method easily result in performance loss due to the pollution of non-object regions residing inside the kernel. In order to better fit the object shape, anisotropic symmetric kernel is introduced with the selection problem existing not only in scale but also extending to orientation. By simultaneously controlling both the scale and orientation, the estimation bias of the kernel can be controlled by the underlying distribution (Fig. 1a), and result in better mode estimation. Nevertheless, objects in practice may have complex shapes that cannot be well described by simple geometric shapes, even when using the most appropriate one (Fig. 1b). With the expectation that the kernel ideally has the shape of the tracked object, some attempts have been made to use asymmetric kernel for dynamic tracking. However, most of them invite constant kernel shape throughout the sequence, few consider to adapt it to the target variation over time.

In this paper, we derive an adaptive data-driven kernel to simultaneously address the kernel scale/orientation selection problem as well as the constancy of the kernel shape in non-rigid object tracking application. Level set technique is novelly introduced to the mean shift sample space to both cope with insufficient low-level information and implement the adaptive kernel evolution and

*Correspondence: dongli@sdu.edu.cn
[2]Shandong University, Weihai, China
Full list of author information is available at the end of the article

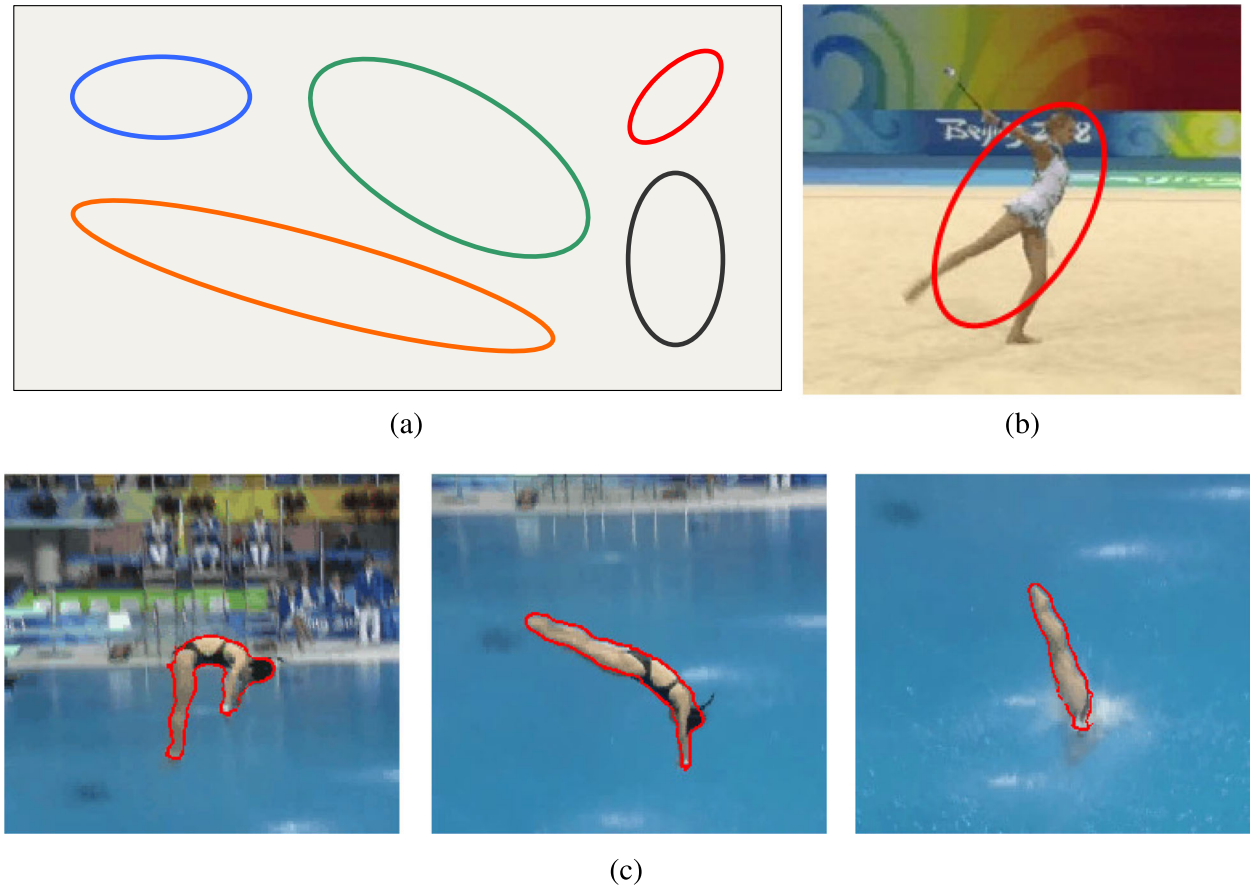

Fig. 1 Motivation and improvement illustration of the proposed method, the frame numbers of **c** are 191, 206, and 219, respectively, in *diving* sequence. **a** Kernel scale/orientation selection. **b** Complex object shape. **c** The proposed data-driven kernel and its adaptation to target variation

update. Since the active contour model is designed to drive the kernel constantly to the direction that maximizes the appearance similarity, the kernel can robustly achieve the adaptation to target variation and act toward the actual target contour simultaneously with the mean shift iterations. As the adaptive kernel continually seizes the target shape, it can give a better estimation bias to produce accurate shift of the mean and successfully avoid the performance loss stemmed from pollution of the non-object regions hiding inside the kernel. Briefly, our main contributions could be summarized as follow:

- In contrast to traditional meanshift methods which use fixed rectangular for target presentation, we introduce the level set model into the meanshift framework to realize non-rigid object contour tracking.
- In contrast to traditional level set method that do not consider any interested target knowledge, we evolve the level set curve in the meanshift sample space to drive the curve whose convergence result maximizes the target appearance similarity.
- We proposed an adaptive data-driven kernel based on the level set model within the meanshift framework, which addresses the fix kernel shape and kernel scale/orientation selection problem facing traditional kernel trackers.

Figure 1 illustrates the motivation and improvement of our proposed method.

2 Related work

2.1 Tracking methods with kernel scale/orientation selection

After the intuitive 10% method in [6], Collins proposed a method [11] using difference of Gaussian mean shift kernel for efficient blobs tracking through scale space. Khan et al. in [12] derive a multi-mode anisotropic mean shift, where the center, size, and orientation of the bounding box are simultaneously estimated during the tracking. In [13], the authors present a probabilistic formulation of kernel-based tracking methods where the EM-estimation conjunction with KL-divergence are used to develop a target-center and kernel bandwidth update scheme.

However, all of them roughly represent the objects by simple geometric shape kernels that easily result in background pollution. In contrast, the proposed data-driven kernel can adapt to the shape of actual object for tracking and as well qualify the samples for appearance model update.

2.2 Tracking methods using asymmetric kernel

In [14], asymmetric kernels are generated using implicit level set functions. After extending the search space to higher dimension, the method simultaneously estimates the new object location, scale, and orientation. Yi et al. propose a method for object tracking based on mean shift algorithm in [15]. They use an object mask to construct the asymmetric kernel and implement probabilistic estimation for the orientation change and scale adaptation. These methods, however, invite constant kernel shape during the tracking task which could not therewith to the object shape in case of out-plane rotations by scale and orientation estimation. In contrast, we evolve the data-driven kernel and adapt it to target variation simultaneously with the mean shift iterations to implement tracking of deformable objects.

2.3 Tracking methods using level set

Level set technique has been widely used for dynamic tracking [16–19]. Bibby et al. [20] derive a posterior framework for robust tracking of multiple previously unseen objects where the shapes are implicit contours represented using level set. In [21], the authors add Mumford-Shah model into the particle filter framework. Once the particle filter gives the candidate positions in prediction step, the level set curve evolution is included, without considering any target bias, to give the candidate contours. In [22], dynamical statistical shape priors are introduced and integrated in a Bayesian framework for level set-based image sequence tracking. In [23], the authors propose a fragments based tracking method within the level set framework, where the whole target and background are segmented by an efficient region-growing procedure. Differently, our method introduce the active contour model to the mean shift sample space to both cope with the insufficient low-level information and obtain the adaptive kernel that maximizes the appearance similarity for non-rigid object tracking within mean shift framework.

3 The mean shift estimation

The mean shift method iteratively computes the closest mode of a sample distribution starting from a hypothesized mode. In specifically, considering a probability density function $f(\mathbf{x})$, given n sample points $\mathbf{x}_i$, $i = 1, \cdots, n$, in d-dimensional space, the kernel density estimation (also known as Parzen window estimate) of function $f(\mathbf{x})$ can be written as

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)}{h^d \sum_{i=1}^{n} w(\mathbf{x}_i)} \tag{1}$$

where $w(\mathbf{x}_i) \geq 0$ is the weight of the sample $\mathbf{x}_i$, and $K(\mathbf{x})$ is a radially symmetric kernel satisfying $\int k(x)dx = 1$. The bandwidth h defines the scale in which the samples are considered for the probability density estimation.

Then, the point with the highest probability density in current scale h can be calculated by mean shift method as follow:

$$m_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)} \tag{2}$$

where the kernel profile $k(x)$ and $g(x)$ have the relationship of $g(x) = -k'(x)$.

The kernel recursively moves from the current location $\mathbf{x}$ to the new location $m_h(\mathbf{x})$ according to mean shift vector and finally, converges to the nearest mode.

4 Methods

4.1 Kernel representation

Kernel is a crucial factor to the performance of the mean shift algorithm, which defines the scale of the target candidate and the number of samples considered in the mode seeking process. Inappropriate kernel may result in either noisy background pollution or poor object localization. An ideal kernel is expected to have the shape of the actual tracked object which may be complex, and with the capability of adapting to the object variation. Level set methods, first proposed by Osher and Sethian in [24, 25], offer a very effective representation of contours and are widely used. The basic idea of the level set approach is to embed the contour C as the zero level set of the graph of a higher dimensional function $\phi(x, y, \tau)$, that is

$$C_\tau = \{(x, y) | \phi(x, y, \tau) = 0\} \tag{3}$$

where τ is an artificial time-marching parameter and then evolve the graph so that this level set moves according to the prescribed flow. In this manner, the level set may develop singularities and change topology, while ϕ itself remains smooth and maintains the form of a graph.

Based on the competitive properties described above, the level set comes into sight as a reasonable consideration of presenting the expected adaptive kernel. A kernel function $K : \mathbb{R}^d \rightarrow \mathbb{R}$ in the mean shift framework is supposed to satisfy

$$K(\mathbf{x}) = k(\|\mathbf{x}\|^2) \tag{4}$$

where $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$ and $k : [0, \infty] \rightarrow \mathbb{R}$ is the profile function with following properties:

- k is non-negative.

- k is non-increasing, i.e., if $a < b$ then $k(a) \geq k(b)$.
- k is piecewise , and $\int_0^\infty k(r)dr < \infty$.

Implicit level set function $\phi(\mathbf{x})$, encoding the signed distances of the pixels $\mathbf{x}$ from the object boundary, provides a smooth and differentiable function, and basically meet the requirements of a mean shift kernel. However, there is an exception that the signed distance function of level set is negative outside the object boundary. Therefore, we truncate the level set function of the outside boundary portion (set to 0) as in [14] and normalize the inside portion to meet the density estimator standard

$$K(x,y,\tau) = \begin{cases} \frac{\phi(x,y,\tau)}{\sum_{\phi(x,y,\tau)>0}\phi(x,y,\tau)}, & \text{if } [x \;\; y]^T \text{ inside } C_\tau \\ 0, & \text{else} \end{cases} \tag{5}$$

Figure 2 illustrates the level set kernel mechanism. In [14], the asymmetric kernel is constructed only for once and

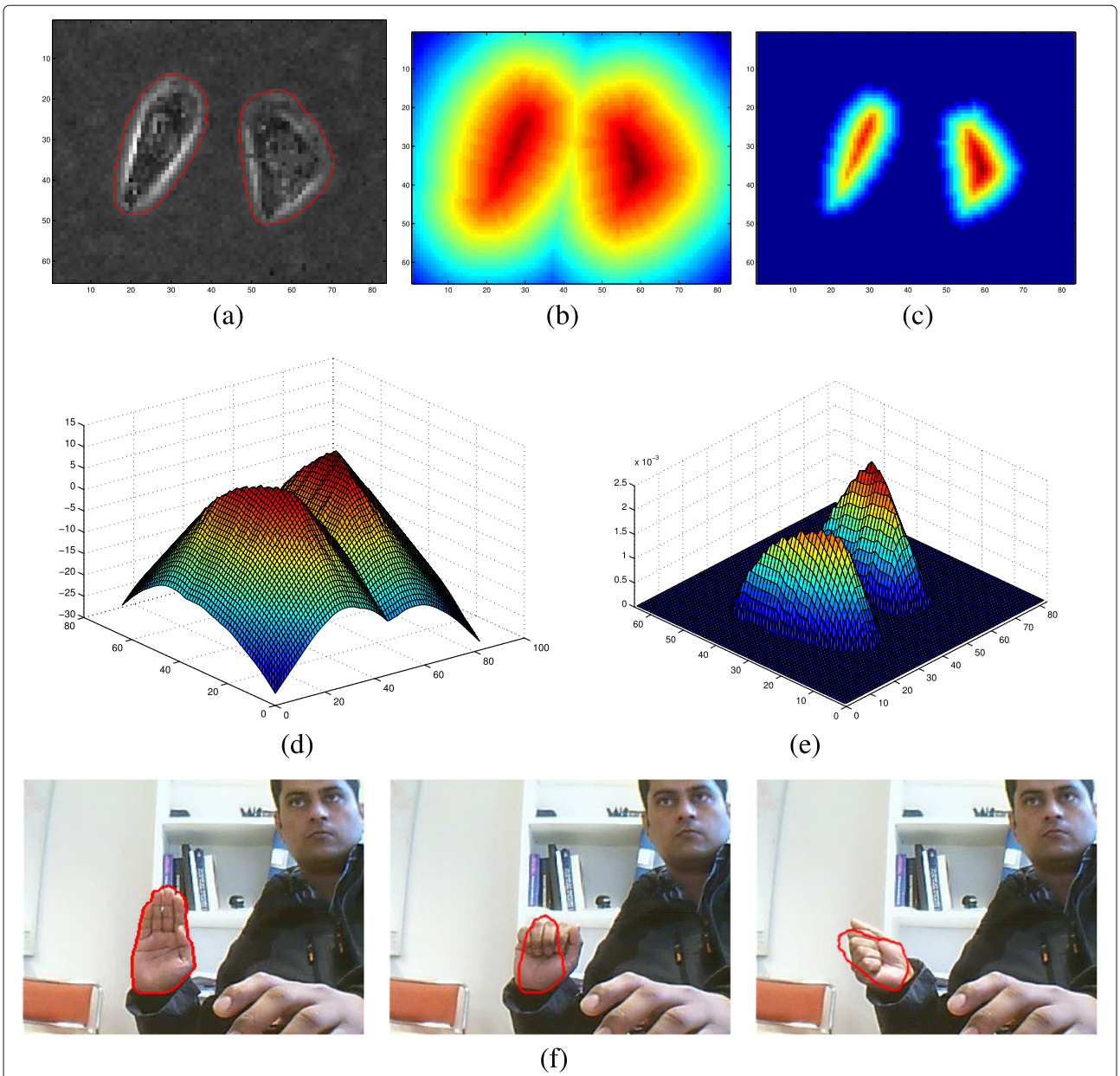

Fig. 2 Illustration of the level set kernel mechanism. **d** and **e** are the corresponding three-dimensional map of the level set function and kernel, the frame numbers of **f** are 0, 87, and 107, respectively, in *hand* sequence. **a** Target contour. **b** Level set function. **c** Level set kernel. **f** Tracking out-plane deforming target with constant kernel shape

used constantly throughout the sequence. Since it does not adapt to object change in shape, the method can only estimate the object scale and orientation of in-plane rotations. In case of 3D or in-depth rotations, it is a challenge for the method to therewith to the object shape (Fig. 2f). Differently, we novelly introduce the active contour model into the mean shift sample space and derive a data-driven kernel, which is able to adapt to the object shape and act toward the actual target contour simultaneously with the mean shift iterations.

4.2 Data-driven kernel evolution

Our goal is to evolve the kernel to the expected image area of the target being tracked. Let $I_\tau : \mathbf{x} \rightarrow \mathbb{R}^m$ denote the image at time τ that maps a pixel $\mathbf{x} = [x \ y]^T \in \mathbb{R}^2$ to a value, where the value is a scalar in the case of a grayscale image ($m = 1$) or a three-element vector for an RGB image ($m = 3$). Effective image preprocessing technical could also be used to generate the value. Let $C(s) = [x(s) \ \ y(s)]^T, s \in [0, 1]$, denote a closed curve in $\mathbb{R}^2$. An implicit function $\phi(x, y)$ is defined as a signed distance function of the curve

$$\phi(x, y, \tau) = \begin{cases} d((x, y), C_\tau), & \text{if } [x \ y]^T \text{ inside } C_\tau \\ 0, & \text{if } [x \ y]^T \text{ at } C_\tau \\ -d((x, y), C_\tau), & \text{if } [x \ y]^T \text{ outside } C_\tau \end{cases} \tag{6}$$

such that the zeroth level set of ϕ is C, that is, $\phi(x, y) = 0$ if and only if $C(s) = [x \ y]^T$ for some $s \in [0, 1]$. Then the contour is deformed in the form of embedding level set function until it minimizes an image-based energy function.

Given an initial kernel region learned from previous observations, we extend the view of candidate object region to a larger ring of neighboring, within which the samples are evaluated by the Bhattacharyya measurement. Therefore, a new kernel function can be adapted without being confined to the current kernel scope. Let q and p denote the color distribution functions generated from the object model and candidate regions, then the weight at pixel $\mathbf{x}$ is given by:

$$w(\mathbf{x}) = \sqrt{q(I(\mathbf{x}))/p(I(\mathbf{x}))} \tag{7}$$

It is obvious that the weight map of the candidate object region contains two kinds of samples. Samples that are more likely to belong to the target than to the background get larger weights, and vice versa for those are more likely derive from the background. In order to distinguish these samples, we include the active contour model into this sample space as an unsupervised clustering manner to automatically separate the samples into two classes (foreground/background) and drive the kernel to the maximum possible area of being the target.

Let m_t and m_b denote the within class weight center of the target and background classes; then, we can define the variance of a cluster D_* around its center by

$$e_*^2 = \sum_{\mathbf{x} \in D_*} \|w(\mathbf{x}) - m_*\|^2 \tag{8}$$

where $* \in \{t, b\}$ denotes the target and background, respectively. Under the intuition that we would like weight values of pixels on the object and background to both be tightly clustered, i.e., low within class variance, we use the sum of squared error criterion as the clustering criterion function

$$J_e = \sum_{* \in \{t,b\}} \sum_{\mathbf{x} \in D_*} \|w(\mathbf{x}) - m_*\|^2 \tag{9}$$

The clustering criterion function optimization is a combinatorial optimization problem and has been proved to be a NP problem. Since the exhaustive computation is unrealistic, we bring this problem into the level set framework and convert the process of iteratively finding approximate solution to the form of level set function evolution. We define the energy function of the active contour as

$$\begin{aligned} E_k(m_*, C) = &\int_{\Omega^+} \|w(\mathbf{x}) - m_t\|^2 \mathrm{d}\mathbf{x} + \int_{\Omega^-} \|w(\mathbf{x}) - m_b\|^2 \mathrm{d}\mathbf{x} \\ &+ \xi \int_C -T(\mathbf{x}) \mathrm{d}\mathbf{x} + \mu \oint_C \mathrm{d}s \end{aligned} \tag{10}$$

where Ω^+ presents the region inside curve C and captures the samples belonging to the object class, while Ω^+ denotes the region outside C and captures the samples of the background class. $T(\mathbf{x})$ is the image gradient for edge detecting

$$T(x, y) = |\nabla [G_\sigma(x, y) * I_\tau(x, y)]|^2 \tag{11}$$

where ∇ denotes spatial gradient operator, $*$ denotes convolution, and G_σ is the Gaussian filter with standard deviation σ. ξ and μ are the coefficients that weight the relative importance of each item.

The first two items are used to measure the within class variation of the object and background classes. The third item is used to ensure the two classes division is on the object boundary. The last item measures the length of the curve C, playing the role of smoothing region boundaries. Therefore, when we minimize the energy function of (10), obviously, we expect to obtain the classification result that both tightly clusters the object/background samples and with division rightly convergent to object edge.

Employing the level set function as a differentiable threshold operator, we unify the integral region and rewrite (10) as

$$E_k(m_*,\phi) = \int_\Omega \|w(\mathbf{x}) - m_t\|^2 H(\phi(\mathbf{x}))\mathrm{d}\mathbf{x} + \int_\Omega \|w(\mathbf{x}) - m_b\|^2 [\,1 - H(\phi(\mathbf{x}))]\,\mathrm{d}\mathbf{x} - \int_\Omega \delta_0(\phi(\mathbf{x}))[\,\xi\, T(\mathbf{x}) - \mu|\nabla\phi(\mathbf{x})|]\,\mathrm{d}\mathbf{x} \quad (12)$$

where $\Omega = \Omega^+ \bigcup \Omega^-$ is the image domain, $H(\cdot)$ denotes the Heaviside function that $H(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{else} \end{cases}$ and $\delta_0(\cdot)$ is the Dirac function.

By fixing the class means of the target and background samples as

$$m_t = \frac{\int_\Omega w(\mathbf{x}) H(\phi(\mathbf{x}))\,\mathrm{d}\mathbf{x}}{\int_\Omega H(\phi(\mathbf{x}))\,\mathrm{d}\mathbf{x}}, m_b = \frac{\int_\Omega w(\mathbf{x})[\,1 - H(\phi(\mathbf{x}))]\,\mathrm{d}\mathbf{x}}{\int_\Omega [\,1 - H(\phi(\mathbf{x}))]\,\mathrm{d}\mathbf{x}} \quad (13)$$

and minimizing the energy functional (12), the associated Euler-Lagrange equation for this functional can be given by

$$0 = \delta_\epsilon(\phi)\left[(w(\mathbf{x}) - m_t)^2 - (w(\mathbf{x}) - m_b)^2 - \xi\, T(\mathbf{x}) - \mu \mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right] \quad (14)$$

and implemented by the following gradient descent:

$$\frac{\partial\phi}{\partial t} = \delta_\epsilon(\phi)\left[-(w(\mathbf{x}) - m_t)^2 + (w(\mathbf{x}) - m_b)^2 + \xi\, T(\mathbf{x}) + \mu \mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right] \quad (15)$$

where div is the divergence operator, and

$$\delta_\epsilon(z) = \frac{1}{\pi}\frac{\epsilon}{\epsilon^2 + z^2} \quad (16)$$

Since the data-driven kernel is designed within the mean shift sample space to act constantly toward the image area with maximum appearance similarity, the kernel curve, in the proposed algorithm, can be steered to the target region from a wide variety of states, without any request of the initial curve that must be inside or outside the target completely.

4.3 Mean shift formulation

For an initial kernel contour $\widehat{C}_{\tau-1}$ learned from previous observations, we evolve it according to the new observation at time τ, I_τ, and the target model q as discussed in Section 4.2. This can be realized by doing a gradient descent on the image energy E_k:

$$\widehat{C}_\tau = evolve(S_\tau, I_\tau, q) = S_\tau^{(M)} \quad (17)$$

where S_τ denotes the curve at time τ, and go through M iterations in the direction of reducing the energy E_k as fast as possible:

$$S^{(\omega)} = S^{(\omega-1)} - \eta^{(\omega)} \nabla_S E_k\left(m_*, S^{(\omega-1)}\right), \quad \omega = 1, 2, \cdots, M \text{ and } S^{(0)} = \widehat{C}_{\tau-1} \quad (18)$$

Based on the object/background division contour $\widehat{C}_\tau$, we can obtain the corresponding kernel $K(x,y)$ as described in Section 4.1. Then the density estimator can be given by

$$\widehat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i\in\Omega^+} K(\mathbf{x} - \mathbf{x}_i) \quad (19)$$

where N is the number of samples in Ω^+, the inside region of $\widehat{C}_\tau$. The mean shift vector that maximizes the density is computed by

$$\Delta\mathbf{x} = \frac{\sum_{i\in\Omega^+} K(\mathbf{x}_i - \mathbf{x}) w(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x})}{\sum_{i\in\Omega^+} K(\mathbf{x}_i - \mathbf{x}) w(\mathbf{x}_i)} \quad (20)$$

Figure 3 illustrates the tracking mechanism of the proposed algorithm.

5 Results and discussion

In this section, firstly, the proposed method was qualitatively evaluated on several video sequences with different challenges for tracking. All the sequences derive from real-world objects records. Then, the proposed method was further tested on two public datasets for quantitative evaluation. In all cases, the target objects and candidates are modeled in RGB space by the weighted histogram with 16 bins along each dimension. The initial curve of the first frame was a rough polygon supplied manually while the subsequent ones were fed by the results of previous frame.

5.1 Qualitative evaluation

The first sequences consist of 230 frames and describe a waving hand with significant shape deformations as well as scaling, rotation changes. From the tracking results shown in Fig. 4 (red), we can see that the proposed method can accurately follow the target due to the adaptation of the data-driven kernel to the object shape variation. For the same sequence, the conventional mean shift tracker (green) could not give well presentation by typical symmetric kernel conjunction with different bandwidths selection.

We further compared three mean shift-based algorithms on a *high jump* sequence to show the superiority of our approach. This sequence records a high jump match, which contains a player undergoing significant shape deformations simultaneously with fast and drastic motion. The three algorithms we tested are (a) standard mean shift using symmetric kernel with different scales selection [6], (b) constant asymmetric kernel-based tracker with both scale and orientation adaptation [14], and (c) the proposed method. Figure 5 shows the tracking results of these

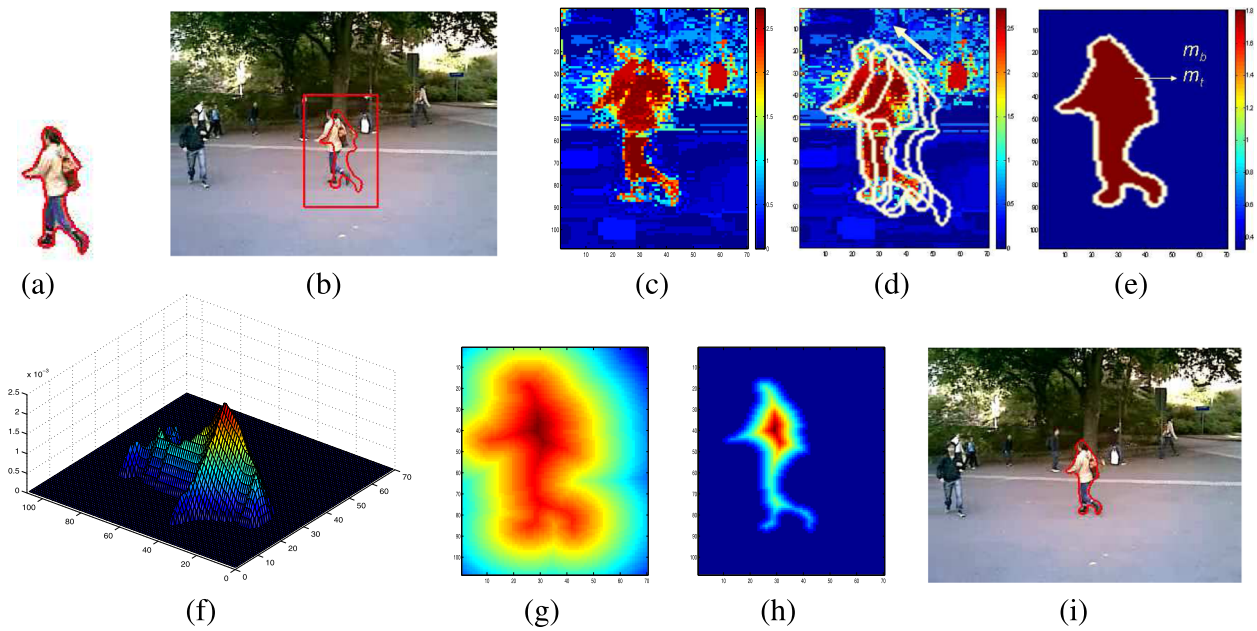

Fig. 3 Tracking mechanism of the proposed method. Employing *pedestrian* sequence as an example, **a** and **b** show the initial curve obtained from previous frame where we admit 15 pixels extension for the new kernel adaptation. **c** The corresponding candidate samples weights. Then, we include the active contour model into this sample space (**d**) to separate them into two classes (foreground/background) and obtain the adaptive kernel that maximizes the appearance similarity (**e**). **g** The level set function of the contour and **f**, **h** the corresponding kernel which is used within the mean shift framework for deforming object tracking tasks (**i**)

algorithms. We can see that in typical mean shift, the pollution of background pixels in the rough kernel region easily results in performance loss and does not guarantee to focus on the target accurately. The algorithm (b), based on constant kernel shape throughout the image frames, is impotent to well present the deforming target only by scale and orientation adjustment. The proposed algorithm, in contrast, effectively adapts the kernel to target variation and obtains pleasant results.

Then, we compared our work with conventional level set-based deformable object tracker [21] on a *pedestrian* sequence. This sequence describes a woman with multi-colored appearance walking in a clutter street with large posture changes and sheltering cases. In [21], the traditional Mumford-Shah method is added within the particle filter framework without considering any target bias. Since the typical level set model emphasizes the intensity consistence only, its convergence on multi-colored region highly depends on the initial curve. Therefore, incompetent results are shown in Fig. 6 a due to the unreliable initial curves derived from the prediction step of particle filter procedure. In contrast, the proposed algorithm, based on the weighted mean shift samples, can simultaneously segment out the two class pixels and obtain the accurate target contour. Additionally, we use the decreasing rate of object size over previous few frames as an occlusion detector. Once detected, we slow down the speed of updating the target density distribution, enabling tracking to resume when the target reappears (Fig. 6b).

Another three challenging sequences were tested to further evaluate the proposed method. The first sequence describes a complex scenario where a girl is moving quickly in a circular path with a boy, undergoing significant scale changes and shape deformation as she moves

Fig. 4 Tracking results of the proposed method (red) and standard mean shift (green) on *hand* sequence. #5, 54, 83, 108, 154

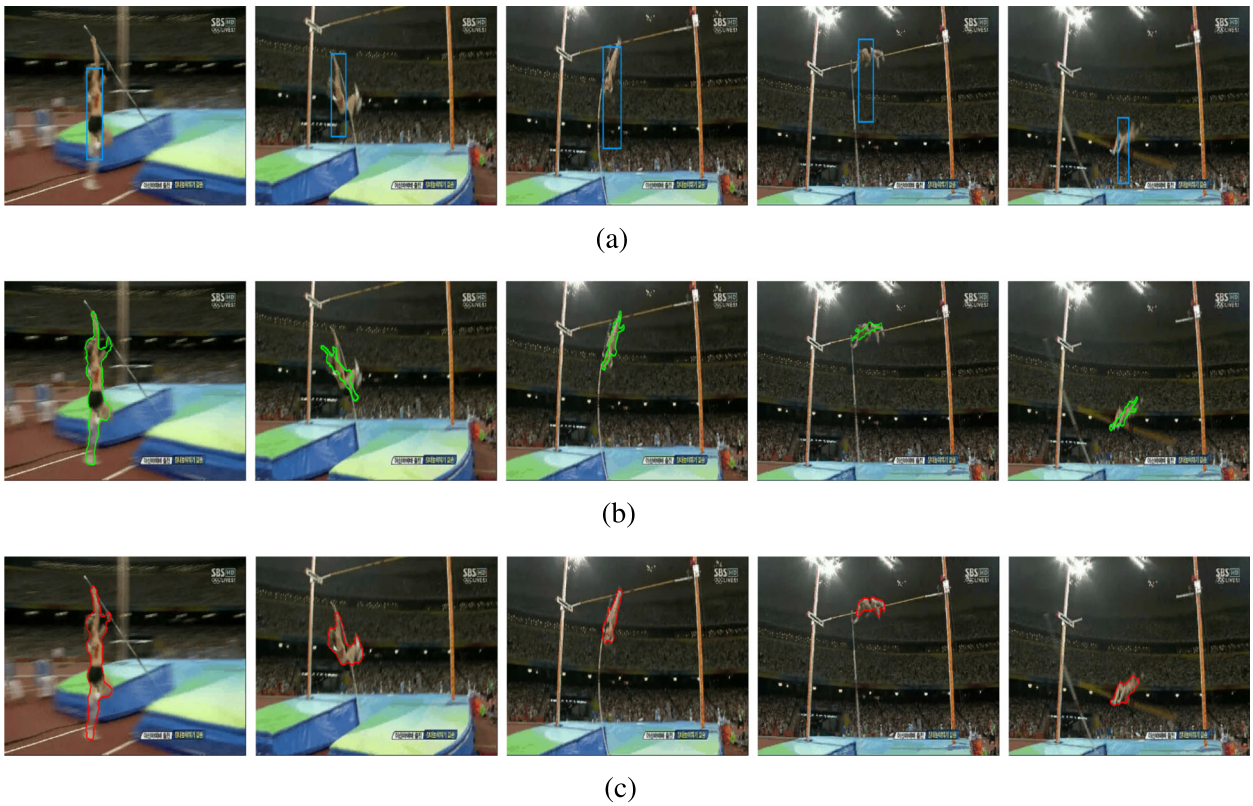

Fig. 5 Tracking results on *high jump* sequence for frames of 0, 13, 27, 39, and 64. **a** Standard mean shift [8]. **b** The method in [34]. **c** The proposed method

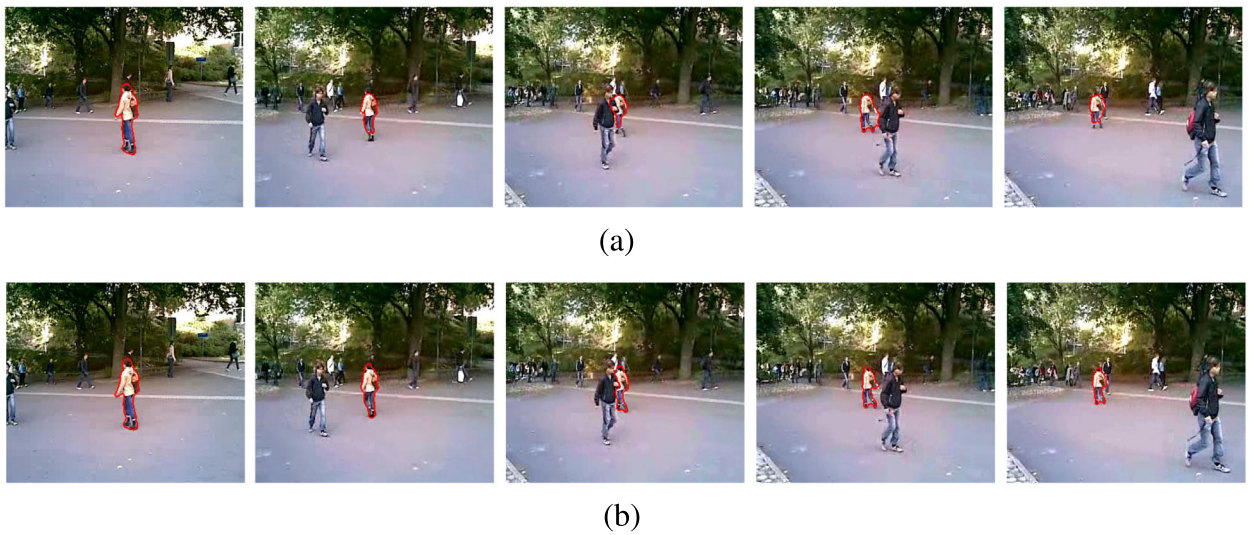

Fig. 6 Tracking results on *pedestrian* sequence for frames of 0, 17, 27, 33, and 47. **a** Conventional level set-based deformable object tracker [26]. **b** The proposed method

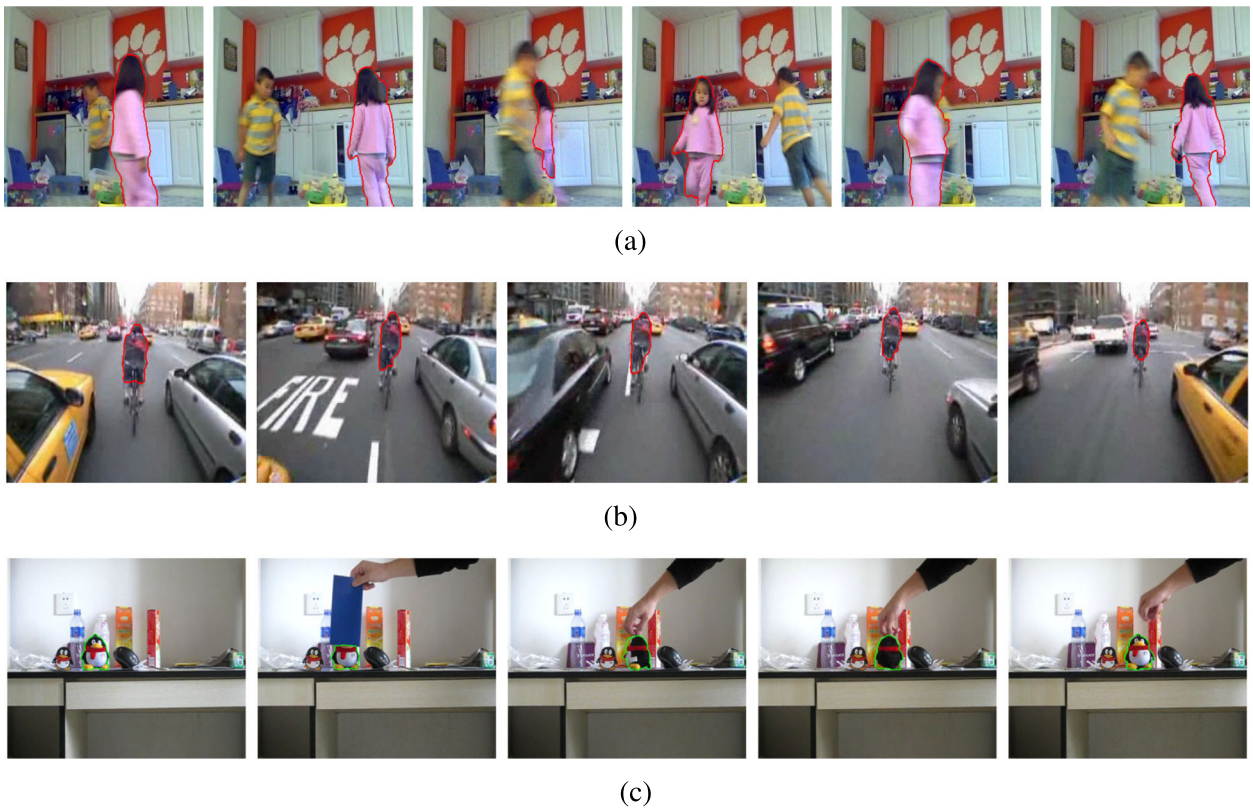

Fig. 7 Experimental results of further evaluation. **a** Tracking results of the proposed method on children sequence for frames of 3, 16, 28, 40, 51, and 64. **b** Tracking results of the proposed method on riding sequence for frames of 100, 216, 239, 268, and 367. **c** Tracking results of the proposed method on toy QQ sequence for frames of 0, 53, 180, 210, and 288

toward or deviating from the camera. It is a challenge for traditional symmetric kernel or intensity edge-based level set methods to represent the child accurately. As we can see in Fig. 7a, the proposed method shows pleasant results, demonstrating the effectiveness of the technical. Compared with the method of [23], where the whole target and background are segmented into intensity consistent fragments and separately modeled in GMM manner, ours include the active contour model in the mean shift sample space and is committed to obtain the adaptive kernel for deformable object tracking within the mean shift framework, overcoming the computational complexity problem facing the traditional contour trackers. The second sequence contains a man riding on a busy road, with the camera moving fast and background changing dramatically. From the tracking results shown in Fig. 7b, we can see that our method performs well even in a complicated scene. The third sequence describes a toy QQ being pulled across the table with clutter background behind and similar icon beside. During this course, large appearance changes occur when the toy is occluded or turned around. Figure 7 c shows the tracking results of this sequence, indicating the competence of the proposed method in dealing with these challenging cases.

5.2 Quantitative evaluation

In this part, for quantitative analysis, we evaluate the proposed method using two public sets of challenging video sequences and compare it to several state-of-the-art tracking methods. The first dataset is VOT2014[1] [26] which comprises 25 sequences (an overall size of more than 10,000 frames), and the second is the VOT2016[2] which consists of 60 sequences. These sequences show various objects with different challenges for visual tracking, including large shape deformations, scale variations, illumination variations, occlusion, and so on.

Firstly, we compare the proposed method with several related bounding box trackers that also make use of the segmentation techniques for target tracking: the DF tracker in [27], which divides the image into several layers that present the probabilities of a pixel taking each feature value to define the distribution field as

[1] http://www.votchallenge.net/vot2014/dataset.html
[2] http://www.votchallenge.net/vot2016/dataset.html

Table 1 Evaluation results of the compared methods on VOT2014 dataset: percentage of correctly tracked frames (score > 0.5)

	Sequence	Pix[28]	DF[27]	HT[32]	SLSM[34]	RPT[29]	Proposed
1	Ball	100	37.31	15.12	100	99.50	100
2	Basketball	41.79	4.00	9.10	37.43	96.55	35.59
3	Bicycle	1.49	98.52	63.43	96.27	83.39	89.3
4	Bolt	10.00	2.57	1.14	2.29	1.43	2.86
5	Car	65.87	39.68	64.68	65.48	100	85.32
6	David	91.69	89.22	72.34	78.57	100	84.94
7	Diving	35.16	27.40	0.46	100	15.98	100
8	Drunk	4.13	18.02	3.14	3.72	100	100
9	Fernando	33.56	62.67	2.05	16.10	65.41	59.93
10	Fish1	1.61	2.29	1.15	6.65	2.29	10.09
11	Fish2	24.19	23.24	5.81	18.06	10.65	9.35
12	Gymnastics	72.46	44.93	9.66	100	42.04	100
13	Hand1	17.43	95.90	100	20.75	21.31	24.59
14	Hand2	19.48	20.60	47.57	48.69	16.85	26.22
15	Jogging	2.28	21.50	80.78	22.15	22.48	100
16	Motocross	6.71	11.59	100	18.29	18.90	15.24
17	Polarbear	100	100	100	100	100	100
18	Skating	9.25	38.00	85.50	53.75	90.00	23.75
19	Sphere	100	9.95	100	100	100	100
20	Sunshade	10.06	50.58	100	68.60	100	100
21	Surfing	98.57	100	100	100	100	100
22	Torus	80.46	20.83	100	100	98.86	100
23	Trellis	87.70	53.08	72.93	39.72	100	36.9
24	Tunnel	1.78	58.55	39.67	25.03	57.73	49.11
25	Woman	17.59	94.47	18.43	88.78	93.80	94.97
	Average	41.3304	44.996	51.7184	56.4132	65.4868	65.9264

image descriptor for target modeling; PixelTrack in [28], which combines a generalized Hough transform based detector with a probabilistic segmentation method in a co-training manner to track deformable objects; and reliable patch tracker in [29], which divides the target into rectangular patches, tracks them with the kernelized correlation filters [30], and integrates them within a particle filter framework [31]. Then, we also compare the proposed method to other relevant contour trackers, which also exploit segmentation technique to extract the target object contour for dynamic tracking. The first method is HoughTrack (HT) proposed by Godec et al. [32], where the authors proposed a patch-based voting algorithm with Hough forests [33]. By back-projecting the patches that voted for the object center, the authors initialize a graph-cut algorithm to segment foreground from background. The second method is the SLSM in [34], in which a single boosting target model is learnt to guide the level set curve evolution to obtain the interested target region.

For the quantitative analysis, for each video, we determine the percentage of frames in which the object is correctly tracked. Since the ground truth annotation included in the datasets is represented by a rotated bounding box, and to let the contour trackers be compared fairly with other bounding box trackers, we measure the tracking accuracy using the Agarwal-criterion [35] as in [32] and [34]. It is defined as score $= \frac{R_T \bigcap R_{\mathrm{GT}}}{R_T}$, where R_T is the output target region from the tracking algorithm and R_{GT} the ground truth. In each image frame, the tracking is considered correct if the Agarwal overlap measure is above a threshold (set to 0.5). Since the VOT2016 dataset

Table 2 Evaluation results of the compared methods on VOT2016 dataset: percentage of correctly tracked frames (score > 0.5)

Methods	Pix [28]	DF [27]	HT [32]	SLSM[34]	RPT [29]	Proposed
Average	40.3302	42.0626	45.7331	47.8365	54.0174	54.5296

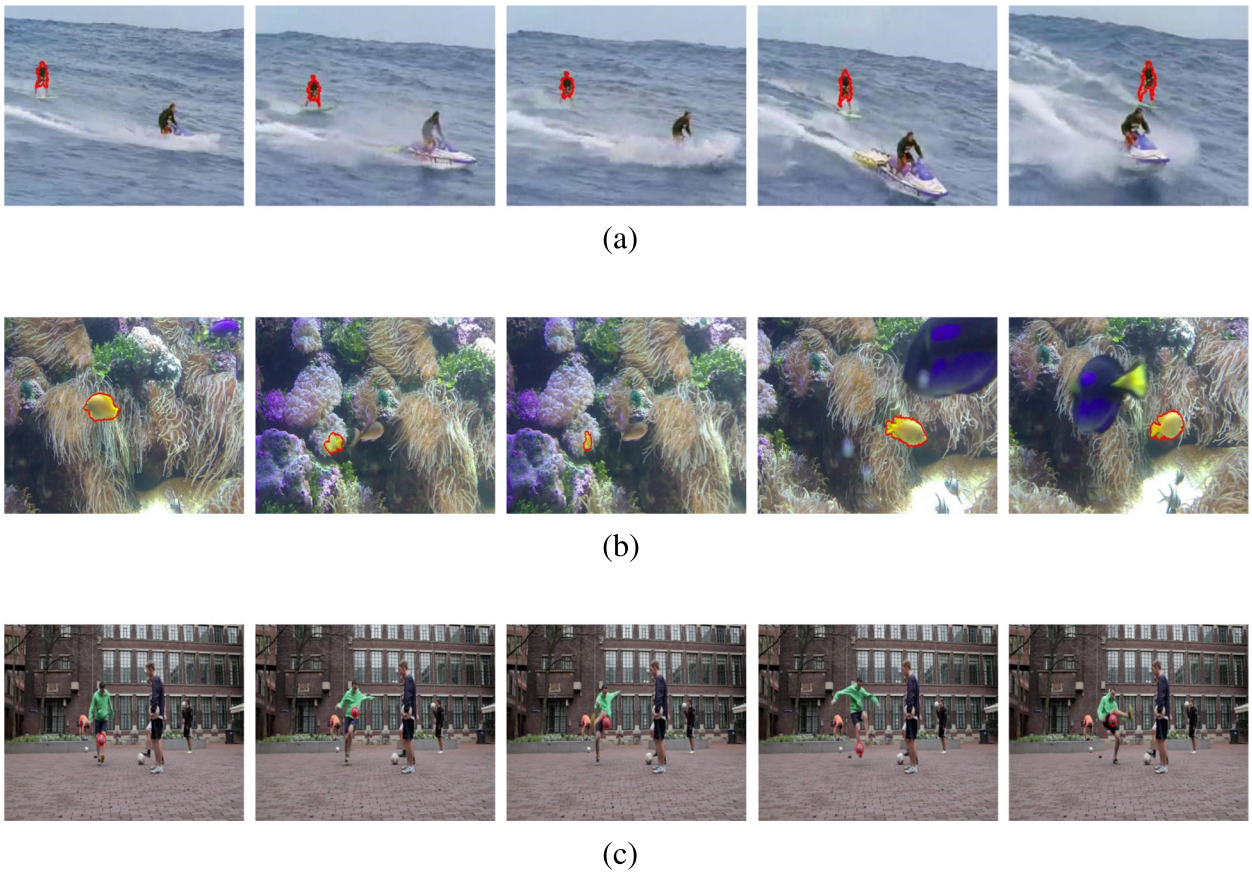

Fig. 8 Visible tracking examples of the proposed method on the VOT2014 and VOT2016 datasets. **a** Tracking results of the proposed method on surfing sequence of the VOT2014 dataset, for frames of 1, 34, 47, 125, and 195. **b** Tracking results of the proposed method on fish3 sequence of the VOT2016 dataset, for frames of 1, 238, 250, 402, and 429. **c** Tracking results of the proposed method on ball1 sequence of the VOT2016 dataset, for frames of 2, 24, 28, 33, and 40

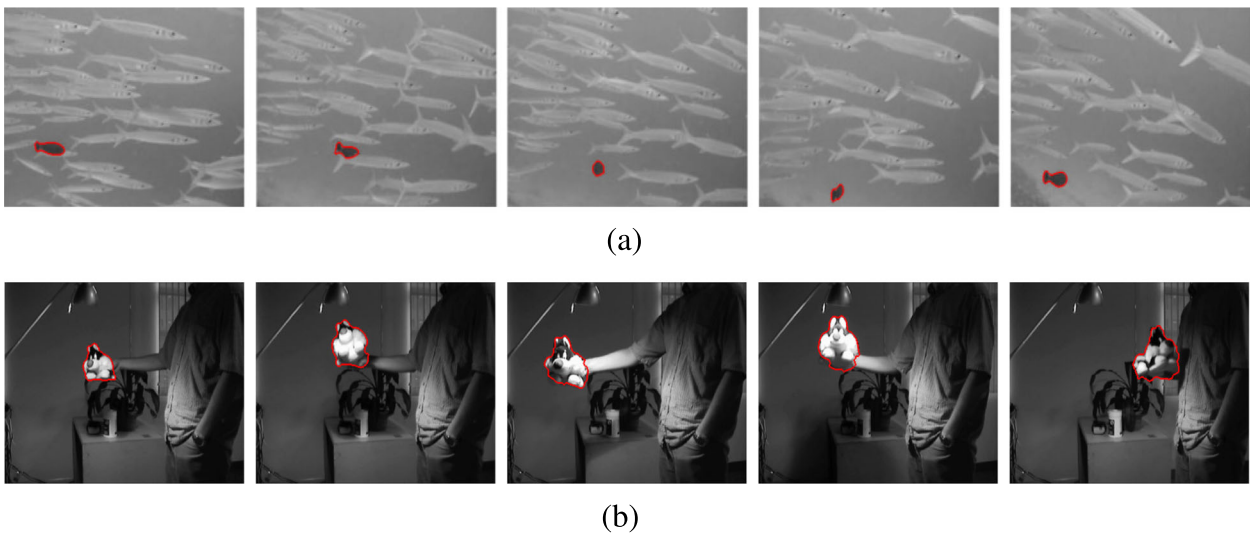

Fig. 9 Experimental results of evaluation on gray scale video sequences. **a** Tracking results of the proposed method on fish sequence for frames of 48, 162, 196, 299, and 355. **b** Tracking results of the proposed method on toy dog sequence for frames of 231, 270, 361, 383, and 542

contains 60 sequences and for the consideration of space, we select the VOT2014 dataset to show the entire evaluation results of the compared methods (see Table 1). As we can see, for 12 out of 25 video sequences the proposed method outperforms the others, and also the average of correct tracking. Table 2 summarizes the quantitative analysis of the compared methods on VOT2016 dataset. Figure 8 gives some visible tracking results of the proposed method on the two datasets.

Finally, we show the ability of the proposed method to work on gray scale images. The first sequence captures a fish whose shape undergoes sudden deformation as it turns or gets occluded. The second sequence describes a toy dog which is held and swayed under a lamp with large appearance and illumination changes as the toy moves and turns. Figure 9 shows the tracking results of these gray scale video sequences. As we can see in images, our work can get pleased performance even with large appearance changes and severe sheltering cases in gray scale images.

6 Conclusion

We have presented a novel data-driven kernel in this paper for non-rigid object tracking. By introducing the active contour model into the mean shift sample space, the adaptive kernel can be evolved and updated to adapt to target variation simultaneously with the mean shift iterations. Since the active contour model is designed to drive the kernel constantly to the direction maximizing the appearance similarity, this adaptive kernel can continually seize the target shape to give a better estimation bias and produce accurate shift of the mean, addressing the problem of constant kernel shape and scale/orientation selection facing typical kernel-based trackers. Experimental results have verified the effectiveness of the proposed method in many complicate scenes.

Acknowledgements
Not applicable.

Authors' contributions
HY and ZB supervise this work. WW and DL partly contribute for the idea discussion, coding of this work, and writing of this manuscript. XS does the main research of this work. All authors read and approved the final manuscript.

Funding
This work is funded by the National Natural Science Foundation of China (No. 61702138 and No. 61602128), and Shandong Province Natural Science Foundation of China (No. ZR2016FQ13), and China Postdoctoral Science Foundation (No. 2019M662360, No. 2017M621275 and No. 2018T110301), and Hong Kong Scholar project of China (No. ALGA4131016116), and Young Scholars Program of Shandong University, Weihai (No. 1050501318006), and Science and Technology Development Plan of Weihai City (No. 1050413421912), and Foundation of Key Laboratory of Urban Land Resources Monitoring and Simulation?Ministry of Land and Resources (No. KF-2019-04-034), and the Fundamental Research Funds for the Central Universities (No. HIT.NSRIF.2019082).

Availability of data and materials
Please contact author for data requests.

Consent for publication
Not applicable.

Competing Interests
The authors declare that they have no competing interests.

Author details
[1]Harbin Institute of Technology, Weihai, China. [2]Shandong University, Weihai, China. [3]Harbin Institute of Technology, Harbin, China.

Received: 17 April 2019 Accepted: 31 January 2020
Published online: 28 February 2020

References

1. N. Wang, W. Zhou, Q. Tian, *et al*, Multi-cue correlation filters for robust visual tracking. IEEE Conf. Comput. Vision Patt. Recogn. (CVPR), 4844–4853 (2018). https://doi.org/10.1109/cvpr.2018.00509
2. G. Zhu, F. Porikli, H. Li, Beyond local search: Tracking objects everywhere with instance-specific proposals. IEEE Confer. Comput. Vision Patt. Recogn. (CVPR), 943–951 (2016). https://doi.org/10.1109/cvpr.2016.108
3. T. Liu, G. Wang, Q. Yang, Real-time part-based visual tracking via adaptive correlation filters. IEEE CVPR, 4902–4912 (2015). https://doi.org/10.1109/cvpr.2015.7299124
4. J. Choi, H. J. Chang, S. Yun, Attentional correlation filter network for adaptive visual tracking. IEEE Confer. Comput. Vision Patt. Recogn. (CVPR) (2017). https://doi.org/10.1109/cvpr.2017.513
5. K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with application in pattern recognition. IEEE Trans. Inf. Theory. **21**(1), 32–40 (1975)
6. D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking. IEEE TPAMI. **25**(5), 564–575 (2003)
7. M. Park, Y. Liu, R. T. Collins, Efficient mean shift belief propagation for vision tracking. IEEE Conf. Comput. Vision Patt. Recogn., 1–8 (2008). https://doi.org/10.1109/cvpr.2008.4587508
8. C. Shen, J. Kim, H. Wang, Generalized kernel-based visual tracking. IEEE Trans. Circuits Syst. Video. Technol. **20**(1), 119–30 (2010)
9. W. Qu, D. Schonfeld, Robust control-based object tracking. IEEE Trans. IP. **17**(9), 1721–6 (2008)
10. C. Yang, R. Duraiswami, L. Davis, Efficient spatial-feature tracking via the mean-shift and a new similarity measure. IEEE Conf. Comput. Vision Patt. Recogn. **1**, 176–183 (2005)
11. R. Collins, Mean-shift blob tracking through scale space. IEEE Conf. CVPR (2003). https://doi.org/10.1109/cvpr.2003.1211475
12. Z. H. Khan, I. Y. Gu, A. G. Backhouse, Robust visual object tracking using multi-mode anisotropic mean shift and particle filters. IEEE Trans. Circ. Syst. Video Technol. **21**(1), 74–78 (2011)
13. Q. A. Nguyen, A. Robles-Kelly, C. Shen, Kernel-based tracking from a probabilistic viewpoint. IEEE Conf. Comput. Vision Patt. Recogn., 1–8 (2007). https://doi.org/10.1109/cvpr.2007.383240
14. A. Yilmaz, Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection. IEEE Conf. CVPR, 1–6 (2007). https://doi.org/10.1109/cvpr.2007.382987
15. K. M. Yi, H. S. Ahn, J. Y. Choi, Orientation and scale invariant mean shift using object mask-based kernel. IEEE Int. Confer. Patt. Recogn., 1–4 (2008). https://doi.org/10.1109/icpr.2008.4761156
16. V. A. Prisacariu, I. Reid, Nonlinear shape manifolds as shape priors in level set segmentation and tracking. IEEE Conf. Comput. Vision Patt. Recogn., 2185–2192 (2011). https://doi.org/10.1109/cvpr.2011.5995687
17. T. Chan, L. Vese, Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
18. J. Lie, M. Lysaker, X. C. Tai, A binary level set model and some applications to mumford-shah image segmentation. IEEE Trans. IP. **15**(5), 1171–1181 (2006)
19. C. Li, C. Xu, C. Gui, M. D. Fox, Level set evolution without re-initialization: A new variational formulation. IEEE Conf. CVPR. **1**, 430–436 (2005)
20. C. Bibby, I. Reid, Real-time tracking of multiple occluding objects using level sets. IEEE Conf. Comput. Vision Patt. Recogn., 1307–1314 (2010). https://doi.org/10.1109/cvpr.2010.5539818

21. Y. Rathi, N. Vaswani, A. Tannenbaum, A. Yezzi, Tracking deforming objects using particle filtering for geometric active contours. IEEE Trans. Patt. Anal. Mach. Intell. **29**(8), 1470–1475 (2007)
22. D. Cremers, Dynamical statistical shape priors for level set-based tracking. IEEE Trans. Patt. Anal. Mach. Intell. **28**(8), 1262–1273 (2006)
23. P. Chockalingam, N. Pradeep, S. Birchfield, Adaptive fragments-based tracking of non-rigid objects using level sets. IEEE ICCV, 1530–1537 (2009). https://doi.org/10.1109/iccv.2009.5459276
24. S. J. Osher, J. A. Sethian, Fronts propagation with curvature dependent speed: Algorithms based on hamilton-jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
25. J. A. Sethian, *Level set methods and fast marching methods*, 2nd edition. (Cambridge University Press, 1999). http://www.cambridge.org/us/catalogue/catalogue.asp?isbn=0521645573
26. S. J. Hadfield, K. Lebeda, R. Bowden, The Visual Object Tracking VOT2014 challenge results (2014)
27. L. Sevilla-Lara, E. Learned-Miller, Distribution fields for tracking. IEEE CVPR, 1910–1917 (2012). https://doi.org/10.1109/cvpr.2012.6247891
28. S. Duffner, C. Garcia, Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects. ICCV, 2480–2487 (2013). https://doi.org/10.1109/iccv.2013.308
29. Y. Li, J. Zhu, S. Hoi, Reliable patch trackers: Robust visual tracking by exploiting reliable patches. IEEE CVPR, 353–361 (2015). https://doi.org/10.1109/cvpr.2015.7298632
30. J. F. Henriques, R. Caseiro, P. Martins, J. Batista, Highspeed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence. **37**(3), 583–596 (2015)
31. A. Doucet, N. Freitas, N. Gordon, Sequential monte carlo methods in practice. Springer-Verlag (2001). https://doi.org/10.1007/978-1-4757-3437-9_4
32. M. Godec, P. M. Roth, H. Bischof, Hough-based tracking of non-rigid objects. ICCV, 81–88 (2011). https://doi.org/10.1109/iccv.2011.6126228
33. J. Gall, A. Yao, N. Razavi, et al., Hough forests for object detection, tracking, and action recognition. IEEE Trans. Patt. Anal. Mach. Intell. **33**(11), 2188–2202 (2011)
34. X. Sun, H. Yao, S. Zhang, D. Li, Non-rigid object contour tracking via a novel supervised level set model. IEEE TIP. **24**(11), 3386–3399 (2015)
35. S. Agarwal, A. Awan, D. Roth, Learning to detect objects in images via a sparse, part-based representation. IEEE TPAMI (2004). https://doi.org/10.1109/tpami.2004.108

Publisher's Note