Non-Rigid Object Tracking via Deformable Patches using Shape-Preserved KCF and Level Sets

Xin Sun[§†] Ngai-Man Cheung[§*] Hongxun Yao[†] Yiluan Guo[§]
Singapore University of Technology and Design[§], Singapore
Harbin Institute of Technology[†], China
sunxintyc@163.com ngaiman_cheung@sutd.edu.sg H.yao@hit.edu.cn guoyl1990@outlook.com

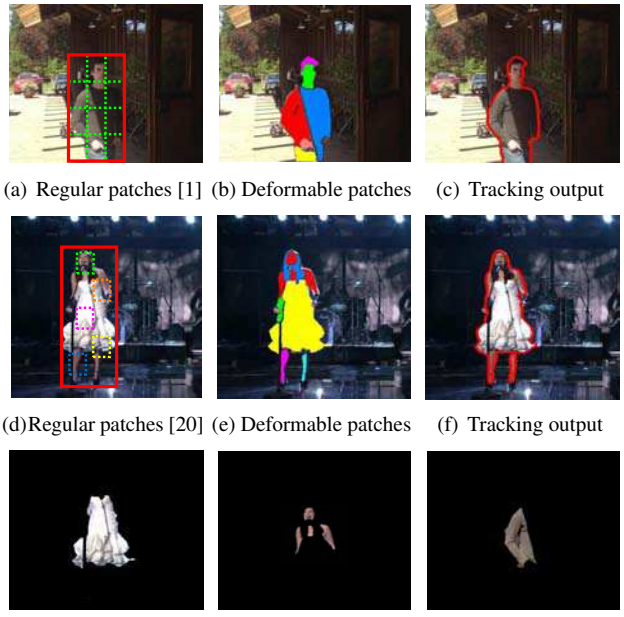
Abstract

Part-based trackers are effective in exploiting local details of the target object for robust tracking. In contrast to most existing part-based methods that divide all kinds of target objects into a number of fixed rectangular patches, in this paper, we propose a novel framework in which a set of deformable patches dynamically collaborate on tracking of non-rigid objects. In particular, we proposed a shape-preserved kernelized correlation filter (SP-KCF) which can accommodate target shape information for robust tracking. The SP-KCF is introduced into the level set framework for dynamic tracking of individual patches. In this manner, our proposed deformable patches are target-dependent, have the capability to assume complex topology, and are deformable to adapt to target variations. As these deformable patches properly capture individual target subregions, we exploit their photometric discrimination and shape variation to reveal the trackability of individual target subregions, which enables the proposed tracker to dynamically take advantage of those subregions with good trackability for target likelihood estimation. Finally the shape information of these deformable patches enables accurate object contours to be computed as the tracking output. Experimental results on the latest public sets of challenging sequences demonstrate the effectiveness of the proposed method.

1. Introduction

Visual tracking refers to the task of generating the trajectories of the moving objects in a sequence of images. It is a fundamental research topic in computer vision and is important in many applications such as surveillance, human-computer interfaces, vision-based control, etc. Despite significant progress over the past decade [9, 24, 32], it remains to be very challenging for tracking in complex scenes due to variations of lighting condition, pose, scale, and view-point over time. In particular, appearance modeling of the target is critical, and it directly affects the robustness of the tracker. A good target model should capture the most distinctive properties between the target object and background, meanwhile with the ability to adapt to target variations. In contrast to describing the entire object with a single global model, dividing the target into patches is effective in capturing local discriminative properties, preserving spatial relationships, and dealing with partial occlusions.

However, most existing part-based methods [28, 15, 31, 1, 20, 18] use a number of fixed rectangular patches to divide the target. They do not adapt to different kinds of ob-



(a) Regular patches [1] (b) Deformable patches (c) Tracking output
(d)Regular patches [20] (e) Deformable patches (f) Tracking output
(g) Non-rigid subregion (h) Complex topology (i) Variation adaption

Figure 1. Comparison of regular part-based trackers (a,d) with the proposed deformable patch tracker (b,c,e,f). (g-i) show some examples of the deformable patches: (g,h) are from the object in (e), which show the capability in capturing non-rigid target subregions with complex topology; (i) is from the object in (b), illustrating its adaptiveness to target variations.

*Corresponding Author: Ngai-Man Cheung

jects and their variations, see Fig.1(a)(d) for examples of two kinds of patch division approaches. Methods using fixed rectangular patches have several limitations. For example, firstly, it is difficult to determine the appropriate patch size for different kinds of objects. Large patches inevitably incur background pollution and also lose the spatial information that resists occlusion, while small patches do not contain enough information for robust tracking and suffer from drift problem within a large flat image region. Second, the rectangular patches provide no mechanism for determining the shape of the object. They only approximately estimate the scale of the bounding box that encloses most rectangular patches to present the tracking results. Thus it is difficult for these trackers to obtain accurate target region for complex non-rigid objects.

To address above limitations, in this paper, we present a novel method that dynamically coordinates a set of deformable patches for non-rigid object tracking. A shape-preserved kernelized correlation filter is proposed within the level set framework for the patches tracking. In contrast to traditional KCF tracker that only emphasizes the location of the rectangle patch region, the proposed SP-KCF takes the shape information of the non-rigid patches into account. By the SP-KCF and level sets manner, the proposed deformable patches have the capability to assume complex topology, and are deformable to adapt to target variations during tracking process (Fig.1). As these patches adaptively capture individual target subregions, we consider their photometric discrimination and shape variation to evaluate the trackability of each individual target subregion, which enable the proposed tracker to take advantage of the subregions with good trackability in the estimation of the target likelihood. Finally, the shape information held by these deformable patches enables high quality results of accurate object contours to be computed as the output of the tracker. **Our specific contributions are:** (i) We propose a novel framework that dynamically coordinates a set of deformable patches for non-rigid object tracking; (ii) We propose a shape-preserved kernelized correlation filter within a level set framework for deformable tracking of the patches; (iii) We propose to determine the trackability of individual target subregions captured by the adaptive patches using photometric discrimination and shape variation, so as to integrate the patches for the estimation of object contours; (iv) We perform experiments with the latest challenging sequences and comparisons with the state-of-the-art.

2. Related work

Tracking by discriminative appearance modeling. Target modeling is the most critical technique in visual tracking. Much effort in the literature has been devoted for learning discriminative target appearance for robust tracking. In [8], the authors map the target and background into multiple feature spaces, then develop an online feature ranking mechanism to select the top-ranked discriminative features for tracking. In [5], the authors propose an online learning method using an incremental linear discriminant analysis for discriminating the appearances between multiple tracked objects. Another popular category [21, 4, 13, 23] is to learn a binary classifier as the implicit appearance model to distinguish the object from its neighboring background. All of these methods describe the target object with a single global model, where the valuable local details and spatial relationship between pixels are ignored.

Part-based trackers. In contrast to global models, Kwon et al. [17] use a patch-based dynamic appearance model in junction with an adaptive Basin Hopping Monte Carlo sampling method to track a non-rigid object. Both Ting et al. [20] and Yang et al. [18] propose the patch-based trackers based on correlation filter and combine the patches within a particle filter framework. However, these methods use fixed rectangular patches and provide no mechanism for determining the shape of the object. In [29], after dividing the target into rectangular patches, the method selects the most discriminative one to build the target model, based on which the active contour procedure is included for extracting the target region. It ignores the contribution of other patches. As the selected rectangular patch can not properly fit the target subregion, its tracking robustness and discrimination evaluation are degraded, which results in the overall tracking performance is sensitive to the patch size.

Segmentation-based methods for dynamic tracking. For accurately extracting the object region, some attempts in literature have been made to use segmenting technique for dynamic tracking. In [12], Godec et al. present a tracking-by-detection approach based on the generalized Hough-transform. They couple the voting based detection and back-projection with a rough segmentation based on GrabCut [25]. Afterwards, Stefan et al. in [10] improve the above HoughTrack to a faster version by using pixel-based descriptors. [30] learns a boosting classifier to model the target and supervise the active contour evolution to obtain the non-rigid target region. However, they use a single global target model, which does not exploit the valuable local properties and spatial relationships of the object parts. [7] employs the GMM models to segment the entire object and background region into fragments in each image frame for dynamic tracking. However, they consider the fragments of equal importance, and do not adjust contributions of individual fragments based on their trackable properties.

3. The deformable patch based tracker

In this section, we describe the proposed deformable patch based method in detail. First, we introduce the patch representation using level sets, then propose the shape-preserved kernelized correlation filter within the level set

framework for deformable tracking of the patches. We evaluate the photometric discrimination and shape variation of the target subregions captured by the deformable patches to measure their tracking reliability, according which the proposed method integrates the tracking results of individual patch trackers for estimation of target object contour.

3.1. Patch representation

Patches are crucial to the performance of part-based trackers. We expect the patches to preserve target shape information, seizing the distinctive information of individual subregions of the tracked object as tracking basis, and meanwhile with the ability of adapting to object variations. Level set methods, first proposed by Osher and Sethian [22, 26, 6], offer a very effective representation of contours. 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 $\phi(\mathbf{x}, \tau)$, that is $C(\tau) = \{\mathbf{x}|\phi(\mathbf{x}, \tau) = 0\}$ 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, we propose to use the level set to represent the expected deformable patches. We implement an essential segmentation procedure on the target region initialized in the first frame of the image sequence to automatically obtain the subregions of the target object. Then in order to prevent over-division and ensure the patch tracking robustness, we group the subregions with similar appearance and close distance into one patch. Therefore, the patches may keep complex topology. Fig.2 shows an example of using the signed distance level set function to represent the deformable patch in Fig.1(h) containing four image subregions.

3.2. Patch tracking via SP-KCF within level set

Let $\mathbf{C}_{t-1} = \{C_{t-1}^1, \cdots, C_{t-1}^{N_{t-1}}\}$ and $\Phi_{t-1} = \{\phi_{t-1}^1, \cdots, \phi_{t-1}^{N_{t-1}}\}$ denote the set of the patch curves at time $t-1$ and their corresponding level set representation, N_{t-1} is the patch number at time $t-1$. The task of tracking the patches $\mathbf{C}_{t-1}$ is to estimate the corresponding patch set $\hat{\mathbf{C}}_t = \{\hat{C}_t^1, \cdots, \hat{C}_t^{N_{t-1}}\}$ and $\hat{\Phi}_t = \{\hat{\phi}_t^1, \cdots, \hat{\phi}_t^{N_{t-1}}\}$ from the new observed image I_t at time t.

Here we propose to use a shape-preserved KCF tracker within the level set framework to track the patches individually. In conventional KCF tracker (readers may refer to [14] for more detials), the classifier is trained in the Fourier domain, using an image patch x centred around the target with size $W \times H$. The KCF considers all cyclic shifts $x_{w,h}, (w,h) \in \{0, ..., W-1\} \times \{0, ..., H-1\}$ as the training examples for the classifier. The expected label of $x_{w,h}$, $y(w,h)$ follow a Gaussian function, which takes a value of

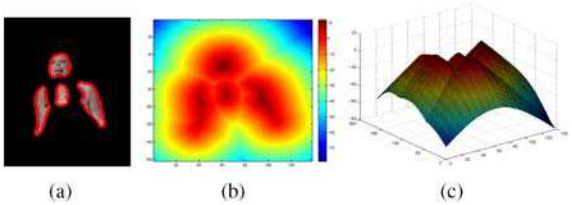


Figure 2. Level set representation of an example patch containing four image subregions. (a) shows the curve of the example patch from Fig.1(h), and (b)(c) show the level set function that represents the patch curve.

1 for a centered target, and smoothly decays to 0 for any other shifts. However, this conventional tracker uses a fixed size of bounding box and only focus on locating the target center. They do not consider the scale, rotation, shape information of the target object, which might lead to drifting when the object changed significantly. Differently, based on the deformable patches, we propose a shape-preserved KCF tracker, that defines the shape-preserved regression targets y to take the non-rigid patch shape into account for robust patch tracking.

Specifically, for each patch C_{t-1}^i in $\mathbf{C}_{t-1}$, an image window of size $W^i \times H^i$ (4 times the size of the minimum outconnected rectangle of C_{t-1}^i) is placed around the patch, which contains all the training examples $x_{w,h}, (w,h) \in \{0, ..., W^i - 1\} \times \{0, ..., H^i - 1\}$. Then, for the regression targets, we calculate the signed distance function of patch curve C_{t-1}^i

$$D(\mathbf{x}) = \begin{cases} d(\mathbf{x}, C_{t-1}^i), & \text{if } \mathbf{x} \text{ inside } C_{t-1}^i \\ 0, & \text{if } \mathbf{x} \text{ at } C_{t-1}^i \\ -d(\mathbf{x}, C_{t-1}^i), & \text{if } \mathbf{x} \text{ outside } C_{t-1}^i \end{cases} \quad (1)$$

where $d(\mathbf{x}, C_{t-1}^i)$ is the Euclidean distance from $\mathbf{x}$ to its nearest point on C_{t-1}^i. Then we normalize the matrix $D(\mathbf{x})$ by mapping its elements to the range of [0 1] to obtain the regression targets $y(w,h)$. Therefore, $y(w,h)$ takes values of 1 at the center ridge of the non-rigid patch, and smoothly decays to 0 along its boundary direction. In this way, different from using the Gaussian function to emphasize only the center position, the proposed tracker accommodates non-rigid shape information of the tracked patch.

Given the training samples $x_{w,h}$ and their corresponding labels $y(w,h)$, the goal of training is to find a function $f^i(z) = w^T z$ that minimizes the squared error over samples $x_{w,h}$ and regression targets $y(w,h)$

$$\min_w \sum_{w,h} |\langle \psi(x_{w,h}), w \rangle - y(w,h)|^2 + \lambda \|w\|^2 \quad (2)$$

where ψ represents the mapping to the Hilbert space induced by the kernel κ. The inner product of x and x' is computed as $\langle \psi(x), \psi(x') \rangle = \kappa(x, x')$. λ is a parameter for the regularization term. After mapping the inputs to a nonlinear feature-space $\psi(x)$, the solution w can be expressed

as $w = \sum_{w,h} \alpha(w,h)\psi(x_{w,h})$. The coefficients vector

$$\alpha = F^{-1}(\frac{F(y)}{F(k^x)+\lambda}) \tag{3}$$

where F and F^{-1} denote the Fourier transform and its inverse, respectively; $k^x = \kappa(x_{w,h}, x_{w,h})$. So far, the target model is learnt by the target appearance $x_{w,h}$ and the transformed classifier coefficients α.

In the new arriving frame I_t, a patch z with the same size of x is cropped out as the search region around the patch region of previous frame. The confidence score is calculated as

$$\hat{f}^i(z) = F^{-1}(F(k^z) \odot F(\alpha)) \tag{4}$$

where $\odot$ is the element-wise product; $k^z = \kappa(z_{w,h}, x_{w,h})$.

We introduce above confidence score of the shape-preserved KCF tracker into the level set framework to evolve the patch curve C^i_{t-1} to the new curve $\hat{C}^i_t$ in frame I_t. We formulate our patch tracking problem as seeking the patch curve $\hat{C}^i_t$ that maximizes the probability

$$p(\hat{C}^i_t|I_t, \hat{f}^i_t) \propto p(\hat{f}^i_t|\hat{C}^i_t)p(I_t|\hat{C}^i_t)p(\hat{C}^i_t) \tag{5}$$

where $p(\hat{f}^i_t|\hat{C}^i_t)$ denotes the likelihood of the observed confidence score enclosed by curve $\hat{C}^i_t$; $p(I_t|\hat{C}^i_t)$ is for photometric segmentation, measuring the photometric consistence of the image region enclosed by $\hat{C}^i_t$; $p(\hat{C}^i_t)$ is for smooth constrain.

Based on above formulation, we define the following level set energy, minimizing which is equivalent to maximizing the probability of (5)

$$\begin{aligned} E(\hat{C}^i_t, m) = &\int_{R^+} -\hat{f}^i_t(\mathbf{x})d\mathbf{x} + \int_{R^-} \hat{f}^i_t(\mathbf{x})d\mathbf{x} \\ &+ \xi \int_{R^+} |I(\mathbf{x}) - m|^2 d\mathbf{x} + \mu\ell(\hat{C}^i_t)\mathbf{x} \end{aligned} \tag{6}$$

where R^+ denotes the image region inside curve $\hat{C}^i_t$ and R^- the region outside the curve; $|I(\mathbf{x}) - m|^2$ measures the photometric consistency of the image region, m is the mean intensity of the region; $\ell(C)$ is the length of the curve; ξ and μ are the coefficients. Therefore, the expected curve $\hat{C}^i_t$ is that encloses a smooth image region with maximum patch confidences.

For the consideration of efficiency, a simple form of two-valued level set function [19] is used to replace the traditional signed distance function, i.e. $\hat{\phi}^i_t(\mathbf{x}, \tau) = 1$ for $\mathbf{x}$ inside $\hat{C}^i_t$ and $\hat{\phi}^i_t(\mathbf{x}, \tau) = -1$ for $\mathbf{x}$ outside $\hat{C}^i_t$. Employing the binary level set function to represent the patch curve $\hat{C}^i_t$ and unify the integral region, the above energy function can be rewritten as

$$\begin{aligned} E(\hat{\phi}^i_t, m) = &\int_\Omega -\frac{1}{2}\hat{f}^i_t(\mathbf{x})(1+\hat{\phi}^i_t) + \frac{1}{2}\hat{f}^i_t(\mathbf{x})(1-\hat{\phi}^i_t) \\ &+ \xi|I(\mathbf{x}) - m|^2(1+\hat{\phi}^i_t) + \mu|\nabla\hat{\phi}^i_t| + \frac{1}{\tau}W(\hat{\phi}^i_t)d\mathbf{x} \end{aligned} \tag{7}$$

where $\Omega = R^+ \bigcup R^-$ is the search region; $m = \frac{\int_\Omega I(\mathbf{x})(1+\phi)d\mathbf{x}}{\int_\Omega (1+\phi)d\mathbf{x}}$; the last item is for binary constraining $\hat{\phi}^{i2}_t = 1$ and W can be defined as $(\hat{\phi}^{i2}_t - 1)^2$. Then the associated Euler-Lagrange equation for this function can be implemented by the following gradient descent:

$$\frac{\partial\hat{\phi}^i_t}{\partial\tau} = \hat{f}^i_t(\mathbf{x}) - \xi|I(\mathbf{x}) - m|^2 + \mu\text{div}(\frac{\nabla\hat{\phi}^i_t}{|\nabla\hat{\phi}^i_t|}) - \frac{1}{\tau}W'(\hat{\phi}^i_t) \tag{8}$$

where div is the divergence operator.

3.3. Patch evaluation

As the deformable patches adaptively capture individual target subregions, we consider their photometric discrimination and shape variation to evaluate their tracking reliability, so as to determine their importance weights in the contribution for the overall tracking task.

We firstly use the augmented variance ratio (AVR) [8], the ratio of the between class variance to the within class variance, to measure the discriminative power of each patch against its local background. For each patch C^i_{t-1} in $\mathbf{C}_{t-1}$, a larger ring of neighboring pixels within a local window are chosen to represent the patch background. By normalizing the image value histograms, we obtain a discrete probability density $p^i(j)$ for the patch C^i_{t-1}, and density $q^i(j)$ for its background. Then the log likelihood of an image value j can be computed by $L^i(j) = \log\frac{\max\{p^i(j),\delta\}}{\max\{q^i(j),\delta\}}$, where δ is a small value that prevents dividing by zero. The log likelihood maps the region into positive for image values associated with the patch region, and negative for values from the background. Then the variance ratio of $L^i(j)$ can be computed as:

$$\text{VR}(L^i; p^i, q^i) = \frac{var(L^i; (p^i + q^i)/2)}{var(L^i; p^i) + var(L^i; q^i)} \tag{9}$$

where $var(L; a) = \sum_j a(j)L^2(j) - [\sum_j a(j)L(j)]^2$ defines the variance of $L(j)$ with respect to a discrete probability density function $a(j)$. The denominator is small when the log likelihood values of pixels in the patch and background classes are tightly clustered, while the numerator is large when the two clusters are widely separated. Therefore, patches with large variance ratio show stronger discriminative power for visual tracking. Fig.3 shows the discrimination values of different patches from the *singer* object shown in Fig.1(e).

Moreover, shape variation of target subregions captured by each individual patch is considered to prevent the error caused by a significant outlier of a false tracked and drift patch. Given the patch set $\mathbf{C}_{t-1}$ and its corresponding tracking output $\hat{\mathbf{C}}_t$, for each patch, we calculate the Hausdorff distance, $D_{\text{H}}(C^i_{t-1}, \hat{C}^i_t)$, of the two corresponding patch curves to measure its shape variation.

Figure 3. Discrimination degrees of different patches from the *singer* object (left). Each bar in the right figure corresponds to the patch with the same color in the middle image.

3.4. Dynamically collaboration

Based on above patch tracking and evaluation, we dynamically coordinate the deformable patches for estimation of target object contour Γ_t. For each tracking result of individual patch tracker, the proposed method weight it by a reliable confidence R_t^i, to dynamically determine its contribution made in the overall tracking task

$$R_t^i = \mathrm{VR}(L^i; p^i, q^i) + \frac{1}{\eta D_{\mathrm{H}}(C_{t-1}^i, \hat{C}_t^i)} \tag{10}$$

where η is the trade off coefficients. This weight is applied on the SP-KCF response $\hat{f}^i(z)$ within the patch region. Through this adaptive weighting, the target likelihood map $\mathbf{R}_t = \sum_{i=1}^{N_{t-1}} R_t^i \hat{f}^i(z)$ puts more emphasis on the discriminative and reliable patches and suppresses the responses of those falsely tracked ones. The object contour Γ_t that encloses maximum target likelihood, i.e. maximizing $p(\Gamma_t|\mathbf{R}_t)$, can be obtained by implementing the level set evolution on map $\mathbf{R}_t$ according to the energy function

$$E(\phi_t^\Gamma) = \int_\Omega -\frac{1}{2}\mathbf{R}_t(\mathbf{x})(1+\phi_t^\Gamma) + \frac{1}{2}\mathbf{R}_t(\mathbf{x})(1-\phi_t^\Gamma) + \mu|\nabla\phi_t^\Gamma| + \frac{1}{\tau}W(\phi_t^\Gamma)\,d\mathbf{x} \tag{11}$$

with the object contour Γ_{t-1} of previous frame as initialization. ϕ_t^Γ is the level set representation of Γ_t. Fig.4 and Algorithm 1 show the tracking framework of the proposed method.

3.5. Model update

In order to accommodate target variations, after obtaining the new target region, we accordingly update the target model, including generating an updated target patch set $\mathbf{C}_t$, and accordingly updating each patch tracker. One option is to segment the new target region to get the new patches. However, this operation increases computational complexity and discards the reference patch distribution from the initial frame. Instead, to keep relatively stable patch distribution as the first reference frame, and balance between adaptiveness to new observations and resistance to accumulated error, we generate the new patch set based on the current set $\hat{\mathbf{C}}_t = \{\hat{C}_t^1, \cdots, \hat{C}_t^{N_{t-1}}\}$. For each patch region of each patch in $\hat{\mathbf{C}}_t$ (note that one patch may consist of several patch regions), we truncate its outside target boundary portion, then group into one patch for the regions whose mean

Algorithm 1 The Deformable Patch Tracker

Require:
The patch set $\mathbf{C}_{t-1}$ and new observed image I_t
Ensure:
The new target contour Γ_t for time t
The updated patch set $\mathbf{C}_t$
1: **for** each patch $C_{t-1}^i \in \mathbf{C}_{t-1}$ **do**
2: Track C_{t-1}^i in I_t with the proposed SP-KCF within the level set framework to obtain the patch curve $\hat{C}_t^i$.
3: **end for**
4: **for** each patch $C_{t-1}^i \in \mathbf{C}_{t-1}$ and $\hat{C}_t^i \in \hat{\mathbf{C}}_t$ **do**
5: Calculate the photometric discrimination and the shape variances to estimate the importance weight R_t^i for patch $\hat{C}_t^i$ by Equ.10
6: **end for**
7: Integrate the tracking results of each individual patch tracker to obtain the new target contour Γ_t by Equ.11
8: Update the patch set $\hat{\mathbf{C}}_t$ to $\mathbf{C}_t$ as well as the corresponding patch trackers
9: **return** Γ_t and $\mathbf{C}_t$.

feature values are within a threshold T_f and spatial distance within threshold T_s. Moreover, for the patch regions belonging to one patch that have large appearance difference or large spatial distance, we separate them into two or more patches. With these operations, we obtain the new patch set $\mathbf{C}_t = \{C_t^1, \cdots, C_t^{N_t}\}$, N_t denotes the updated patch number of time t. Then for each new patch C_t^i in $\mathbf{C}_t$, we update its corresponding SP-KCF patch tracker using the new observation of the patch with the learning rate proportional to the patch weight. Then tracking repeats for time $t+1$.

4. Experimental results

In this section, we evaluate the proposed method using two latest public sets of challenging video sequences, and compare it to several state-of-the-art tracking methods. The first dataset is VOT2014[1] [16] which comprises 25 sequences (an overall size of more than 10,000 frames) and the second is the new released VOT2016[2] (same as VOT2015) 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. The parameters are set as following: $\lambda = 1e-4$; $\xi = 0.5 \times 255^{-2}$; $\mu = 0.15$; and $\eta = 0.03$. The neighborhood for calculating the variance ratio is selected as the region that surrounds the patch and within the rectangle of two times the size of the minimum out-connected rectangle of the patch. The initial curve of a target in the first frame was a manually drawn polygon that simulates the target contour

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

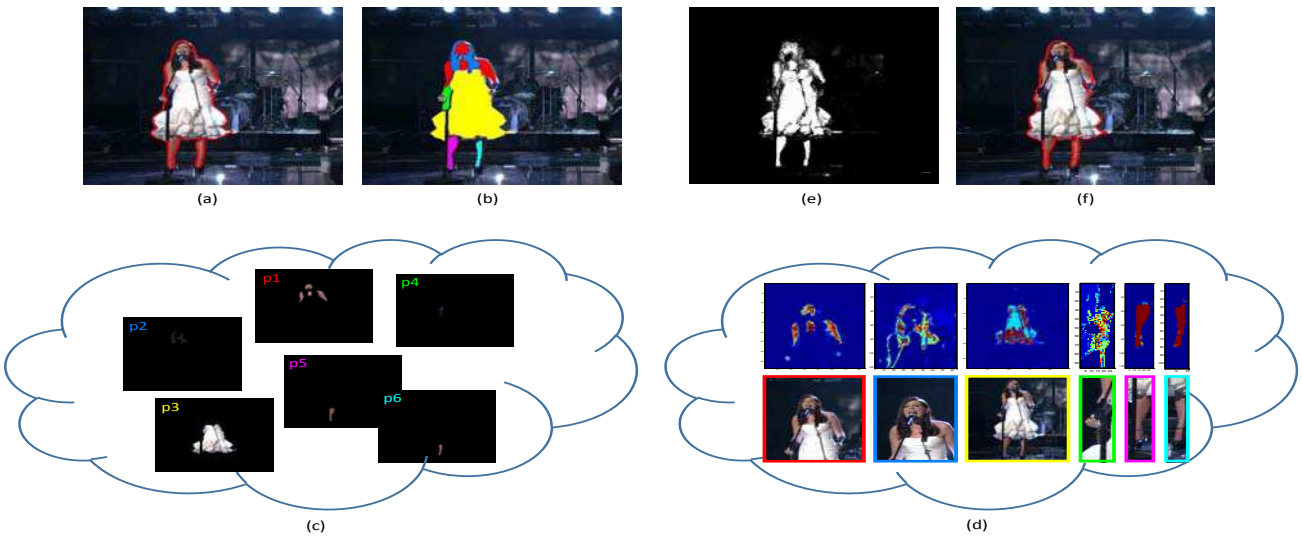


Figure 4. Tracking framework of the proposed method. (a) shows the target region derived from the previous tracking result of time $t-1$; (b)(c) shows the deformable patch set of the target. These patches are tracked individually by a SP-KCF and level set manner in the new arrived image I_t. By exploiting the photometric discrimination and shape variation of the individual target subregions, the proposed method estimate the patch likelihood (d) by $R_t^i \hat{f}^i(z)$ within each tracked patch region, and dynamically coordinates them to obtain the joint target likelihood (e), based on which the level set evolution returns the object contour (f) as tracking output of time t.

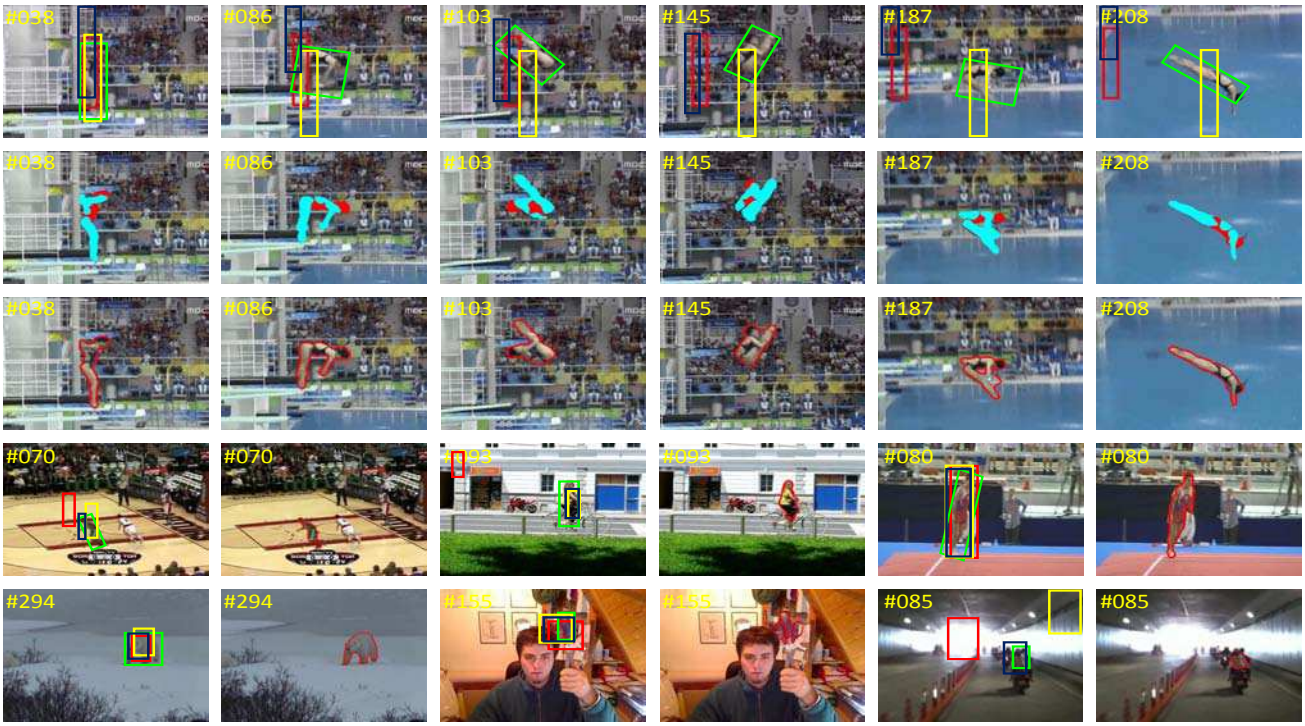


Figure 5. Comparison results of the proposed method with regular bounding box trackers: yellow: the DF tracker [27]; red: the PixelTrack [10]; blue: the reliable patch tracker [18]; green: the groundtruth. The second row shows the examples of the proposed deformable patches, which enable the proposed tracker to extract the accurate target contour as tracking output, shown by red contours in the images.

while the subsequent ones were fed by the result of previous frame. Grabcut [25] can also be used as an alternative to produce a good quality initial contour from a bounding box annotation. We perform region-growing to segment the initial target region. After discarding small fragments, we group the local regions with similar mean value and within a distance of 15 pixels into one patch, then select up to 6 patches of larger size as the initial patch set $\mathbf{C}_0$.

4.1. Comparisons with bounding box trackers

Firstly, we compare the proposed method with several related bounding box trackers that also make use of the division or segmentation techniques for the tracked target modeling: a) 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 image descriptor for target modeling. b) PixelTrack in [10], which combines a generalised Hough transform based detector with a probabilistic segmentation method in a co-training manner to track deformable objects; c) reliable patch tracker in [18], which divides the target into rectangular patches and tracks them with the kernelized correlation filters [14], then integrates them within a particle filter framework [2]. Fig.5 shows the tracking results of these compared methods. From the *diving* sequence shown in the upper row, we can see it is a challenge for the bounding box trackers to accurately locate the target during it undergoes dramatic shape deformation. As the bounding box inevitably introduce a lot of background pixels, the established target model can not provide accurate information to better distinguish between the target and background, which result in deviation from the ground truth. Similar phenomenons of tracking drift or poor target location can be observed in other more targets, lower rows in Fig.5. In contrast, the proposed algorithm, dynamically coordinating the deformable patches (second row), extracts the accurate contours to describe the target as well as qualified samples to establish the appearance model.

4.2. Comparisons with dynamic contour trackers

In this section, we 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. [12], where the authors proposed a patch-based voting algorithm with Hough forests [11]. By back-projecting the patches that voted for the object centre, the authors initialise a graph-cut algorithm to segment foreground from background. The second method is the SLSM in [30], in which a single boosting target model is learnt to guide the level set curve evolution to obtain the interested target region. Fig.6 presents some tracking results of the three methods on *Jogging* sequence. Both the HoughTrack and the SLSM tracker lose the target after frame 85 when the target passed the street lamp and occlusion occurred, since they use a single global model to represent the target which is easily polluted by the similar background. Unlikely, the proposed method tracks the individual patches within their own local background and evaluate their trackability to emphasis the contributions of the discriminative and reliable patches made in the overall target likelihood estimation, thus can achieve robust tracking. Moreover, we accordingly use the patches' trackability to adjust the learning rate of the corresponding patch trackers, enables the proposed tracker to slow down the update speed and keep conservative to deal with abnormal situations, allowing tracking to resume when the target reappears.

	Sequence	Pix[10]	DF[27]	HT[12]	SLSM[30]	RPT[18]	Proposed
1	ball	**100**	37.31	15.12	**100**	99.50	**100**
2	basketball	41.79	4.00	9.10	37.43	96.55	**97.24**
3	bicycle	1.49	**98.52**	63.43	96.27	83.39	87.82
4	bolt	**10.00**	2.57	1.14	2.29	1.43	3.43
5	car	65.87	39.68	64.68	65.48	**100**	87.30
6	david	91.69	89.22	72.34	78.57	**100**	78.57
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**	64.04
10	fish1	1.61	2.29	1.15	6.65	2.29	**21.79**
11	fish2	**24.19**	23.24	5.81	18.06	10.65	13.55
12	gymnastics	72.46	44.93	9.66	**100**	42.04	**100**
13	hand1	17.43	95.90	**100**	20.75	21.31	16.80
14	hand2	19.48	20.60	47.57	**48.69**	16.85	17.23
15	jogging	2.28	21.50	80.78	22.15	22.48	**100**
16	motocross	6.71	11.59	**100**	18.29	18.90	12.20
17	polarbear	**100**	**100**	**100**	**100**	**100**	**100**
18	skating	9.25	38.00	85.50	53.75	**90.00**	88.75
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**	39.72
24	tunnel	1.78	**58.55**	39.67	25.03	57.73	51.98
25	woman	17.59	94.47	18.43	88.78	93.80	**96.65**
	average	41.3304	44.996	51.7184	56.4132	65.4868	**71.0828**

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

	Methods	Pix[10]	DF[27]	HT[12]	SLSM[30]	RPT[18]	Proposed
	average	40.3302	42.0626	45.7331	47.8365	54.0174	**56.7567**

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

For the quantitiative 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 [3] as in [12] and [30]. It is defined as $score = \frac{R_T \bigcap R_{GT}}{R_T}$, 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 contains 60 sequences and for the consideration of space, we select the VOT2014 dataset

Figure 6. Comparison results of the proposed method with other dynamic contour trackers: upper row: the HoughTrack (HT) [12]; middle row: the SLSM [30]; bottom row: the proposed deformable patch tracker.

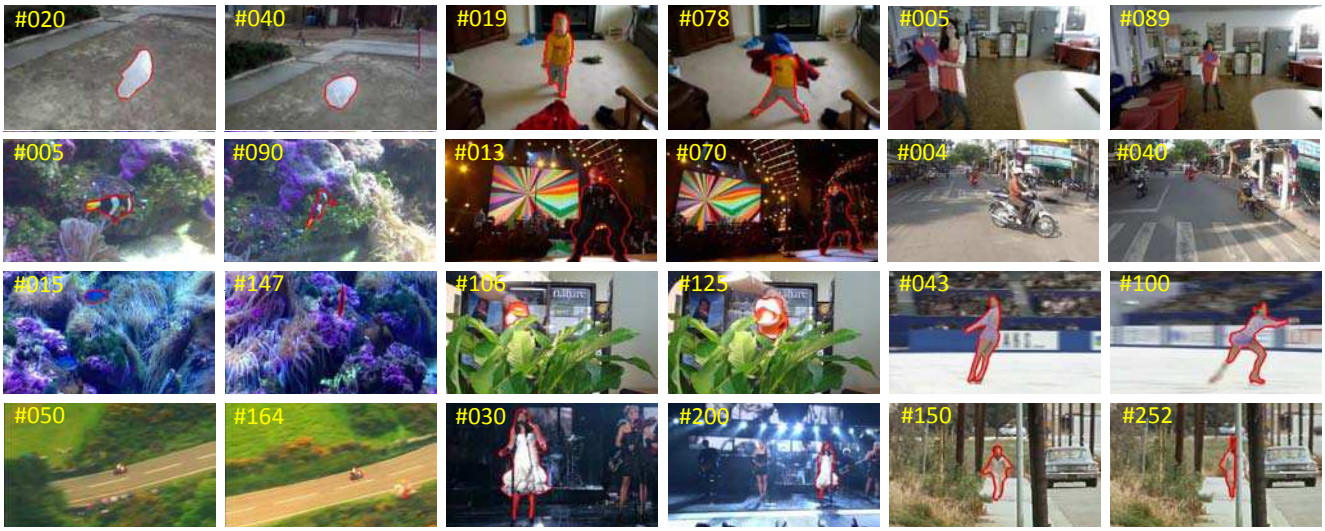


Figure 7. Tracking results of the proposed method on the VOT2016 dataset.

to show the entire evaluation results of the compared methods (see Table I). As we can see, for 13 out of 25 video sequences the proposed method outperforms the others, and also the average of correct tracking. Table II summarises the quantitiative analysis of the compared methods on VOT2016 dataset and Fig.7 gives some visible tracking results of the proposed method on the VOT2016 dataset.

5. Conclusion

We have presented a novel framework in which a set of deformable patches dynamically collaborate on tracking of non-rigid objects. By applying a shape-preserved kernelized correlation filter within the level set framework for deformable tracking, the proposed patches can assume complex topology and are adaptive to target variations. Furthermore, we exploit the photometric discrimination and shape variation of each captured target subregion to evaluate the tracking reliability of each patch tracker, which enables the proposed system to dynamically take advantage of those subregions with good trackability for target likelihood estimation. Shape information held by these deformable patches enables accurate object contours to be computed as the tracking output. Experimental results on latest public sets of challenging sequences verified the effectiveness of the proposed method.

6. Acknowledgement

This research is supported in part by the National Research Foundation Singapore under its Interactive Digital Media (IDM) Strategic Research Programme, in part by the National Natural Science Foundation (Grant No. 61602128 and 61472103) and Shandong Province Natural Science Foundation (Grant No. ZR2016FQ13) of China. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors only.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[2] A.Doucet, N. Freitas, and N.Gordon. Sequential monte carlo methods in practice. *Springer-Verlag*, 2001.

[3] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE TPAMI*, 2004.

[4] S. Avidan. Support vector tracking. *IEEE TPAMI*, 26(8):1064–1072, 2004.

[5] S. Bae and K. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. *IEEE CVPR*, pages 1218–1225, 2014.

[6] T. Chan and L. Vese. Active contours without edges. *IEEE Trans. on IP*, 10(2):266–277, 2001.

[7] P. Chockalingam, N. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. *IEEE ICCV*, pages 1530–1537, 2009.

[8] R. Collins, L. Yanxi, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE TPAMI*, 27(10):1631–1643, 2005.

[9] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] S. Duffner and C. Garcia. Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects. *ICCV*, pages 2480–2487, 2013.

[11] J. Gall, A. Yao, N. Razavi, and et al. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011.

[12] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. *ICCV*, pages 81–88, 2011.

[13] H. Grabner and H. Bischof. On-line boosting and vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 260–267, 2006.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE TPAMI*, 2015.

[15] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1822–1829, 2012.

[16] M. Kristan, R. Pugfelder, A. Leonardis, and et al. The visual object tracking vot2014 challenge results. *ECCV Workshop on Visual Object Tracking Challenge*, 2014.

[17] J. Kwon and K. M. Lee. Tracking of a non-rigid object via a patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. *IEEE CVPR*, 2009.

[18] Y. Li, J. Zhu, and S. Hoi. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. *IEEE CVPR*, pages 353–361, 2015.

[19] J. Lie, M. Lysaker, and X. C. Tai. A binary level set model and some applications to mumford-shah image segmentation. *IEEE Trans. on IP*, 15(5):1171–1181, 2006.

[20] T. Liu, G. Wang, and Q. Yang. Real-time part-based visual tracking via adaptive correlation filters. *IEEE CVPR*, pages 4902–4912, 2015.

[21] S. Lucey. Enforcing non-positive weights for stable support vector tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[22] S. J. Osher and J. A. Sethian. Fronts propagation with curvature dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[23] T. Parag, F. Porikli, and A. Elgammal. Boosting adaptive linear weak classifiers for online learning and tracking. *IEEE CVPR*, 2008.

[24] Y. Qi, S. Zhang, L. Qin, and et al. Hedged deep tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4303–4311, 2016.

[25] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. *ACM SIGGRAPH*, pages 309–314, 2004.

[26] J. A. Sethian. Level set methods and fast marching methods, 1999. Cambridge University Press, 2nd edition.

[27] L. Sevilla-Lara and E. Learned-Miller. Distribution fields for tracking. *IEEE CVPR*, pages 1910–1917, 2012.

[28] G. Shu, A. Dehghan, O. Oreifej, and et al. Part-based multiple-person tracking with partial occlusion handling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1815–1821, 2012.

[29] X. Sun, H. Yao, and S. Zhang. Contour tracking via on-line discriminative appearance modeling based level sets. *IEEE International Conference on Image Processing*, pages 2365–2368, 2011.

[30] X. Sun, H. Yao, S. Zhang, and D. Li. Non-rigid object contour tracking via a novel supervised level set model. *IEEE TIP*, 24(11):3386–3399, 2015.

[31] B. Yang and R. Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. *ECCV*, pages 484–498, 2012.

[32] G. Zhu, F. Porikli, and H. Li. Beyond local search: Tracking objects everywhere with instance-specific proposals. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 943–951, 2016.