A Novel Supervised Level Set Method for Non-Rigid Object Tracking

Xin Sun, Hongxun Yao, Shengping Zhang
Harbin Institute of Technology, 92 West Dazhi Street, Harbin 150001, China
sunxintyc@163.com, {hongxun.yao,shengping.zhang}@gmail.com

Abstract

We present a novel approach to non-rigid object tracking based on a supervised level set model (SLSM). In contrast with conventional level set models, which emphasize the intensity consistency only and consider no priors, the curve evolution of the proposed SLSM is object-oriented and supervised by the specific knowledge of the target we want to track. Therefore, the SLSM can ensure a more accurate convergence to the target in tracking applications. In particular, we firstly construct the appearance model for the target in an on-line boosting manner due to its strong discriminative power between objects and background. Then the probability of the contour is modeled by considering both the region and edge cues in a Bayesian manner, leading the curve converge to the candidate region with maximum likelihood of being the target. Finally, accurate target region qualifies the samples fed the boosting procedure as well as the target model prepared for the next time step. Positive decrease rate is used to adjust the learning pace over time, enabling tracking to continue under partial and total occlusion. Experimental results on a number of challenging sequences validate the effectiveness of the technique.

1. Introduction

Object tracking, which refers to the task of generating the trajectories of the moving objects in a sequence of images, is a challenging research topic in the field of computer vision. Real-world objects are typically difficult to track due to unstable lighting condition, pose variations, scale changes, view-point changes, and camera noise, etc. Many fixed target model based tracking algorithms [6, 7, 16] so are unable to track over long time intervals. To deal with this problem, many tracking algorithms adapt to the appearance changes of the target objects by on-line update, among which the boosting approaches show impressive performances and have received a considerable amount of attention. In contrast with distinguishing the target from background by constructing two separate models respectively, boosting approaches are more inclined to seize properties of most discriminative between them.

However, most boosting based algorithms use rectangle or oval to represent the tracking results while objects, in practice, may have complex shapes, for example, hands, head, and shoulders that cannot be well described by simple geometric shapes. Some attempts in literature have been made to use silhouette or contour, segmenting technique for dynamic tracking [21, 13, 4, 10]. Level set technique [14, 20, 3, 12], as an implicit representation of contours and able to deal with changes in topology, is widely used. The contour is represented as the zero level set of the graph of a higher dimensional function and deformed until it minimizes an image-based energy function. Binary level set model, proposed in [12], uses a two-valued level set function to replace the signed distance function of traditional Chan-Vese manner [3] and greatly improves the computational efficiency. From a performance perspective, the binary level set model is more inclined to segment out the region with consistent intensity, which is similar to the thresholding segmentation method. Recently, many researchers apply level set models to visual tracking. However, few refine them by the prior target knowledge and deal with the problem of multi-mode target segmentation.
In this paper, we present a novel supervised level set model (SLSM) for non-rigid object tracking. Instead of acting towards intensity consistency direction, the curve evolution process of the SLSM is target-oriented and supervised by the specific target model of tracking application. Boosting approach is invited for online construction of the target appearance model due to its strong discriminative power between objects and background. Under the guidance of the prior target knowledge, the proposed SLSM can achieve the multi-mode target segmentation, and the curve finally converges to the candidate region with maximum likelihood of being the target. We use the positive decrease rate to adjust the target learning pace over time, which enables tracking to continue under partial and total occlusion. Fig. 1 shows some tracking examples of our method in various challenging cases.

2. Relate work

Tracking methods with online appearance learning: Han et al. [9] robustly deal with the variations of lighting condition, pose, scale, and view-point over time by approximately estimating the pixel-wise color density in a sequential manner. Ross et al. [19] use the incremental principal component analysis to adaptively track object with large changes in pose, scale, and illumination. The work of Grabner et al. [8] and Parag [15] show impressive results of using classifier as implicit appearance model. They initially learn a binary classifier to distinguish the object of interest from the (neighboring) background and then apply it in each new frame to locate the position of the object. All of these tracking methods, however, use simple geometric shapes to represent the objects, where the background pixels in the target region may pollute the subsequent updating results. In contrast, our method utilizes the accurate contour to represent the target as well as qualified samples for updating the appearance model.

Tracking methods for non-rigid objects: Zhao et al. [22] focus on the humans shape modeling and successfully tracks humans in crowded environments where occlusion persistently occurs. Ramanan et al. [17] firstly construct an appearance model for the target which is used afterward as detecting model for tracking. Good results have been obtained on articulated persons. These two methods, however, basically assume that specific models of the targets are given. Kwon et al. [11] use a patch-based dynamic appearance model in junction with an adaptive Basin Hopping Monte Carlo sampling method to successfully track a non-rigid object, where no prior knowledge of the specific target model is requested, neither does the off-line training phase. Similarly, our method explicitly tackles the target change by on-line appearance learning.

Close work on level set based tracking methods: Bibby et al. [2] derive a probabilistic framework for robust tracking of multiple previously unseen objects. The observed image data is used to compute a posterior over the objects poses, shapes and relative depths where the shapes are implicit contours represented using level-sets. In [18], 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, with no prior target knowledge, to give the candidate contours. In contrast, our method supervises the curve acting orientation by the current learnt target model.

3. Curve evolution using level sets

Level set methods, first proposed by Osher and Sethian in [14, 20], offer a very effective implementation of curve evolution. 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 $\phi(x, y, k)$, that is

$$C_k = \{(x, y)|\phi(x, y, k) = 0\}$$

where $k$ 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 $\phi$ itself remains smooth and maintains the form of a graph.

Generally, the curve evolution equation can be defined as

$$\frac{\partial C}{\partial k} = VN$$

where $V$ represents the speed of curve evolution while $N$ represents the inward unit normal vector. Then we can get the following equation:

$$\frac{\partial \phi}{\partial k} + \nabla \phi \cdot \frac{\partial C}{\partial k} = 0$$

Based on the definition of the level set function $\phi(x, y, k)$ described above, the vector $N$ can be written as $N = -\nabla \phi/||\nabla \phi||$. Then we can obtain the level set implementation corresponding to the curve evolution equation (2):

$$\frac{\partial \phi}{\partial k} = V||\nabla \phi||$$

Given an initial curve, one must generate an initial level set function. Further more, the level set function also needs to be re-initialized continually during its update process which usually takes a lot of computing time.

4. The supervised level set model

In this section, we analysis the general curve acting principle of the level set model which gives the antecedent of our improvement. Then we describe the proposed supervised level set model in detail, which is competent to deal with real object tracking problems.
4.1. Curve acting principle

In binary level set model [12], a piecewise constant-valued function \( u \) is used to approximate the intensity distribution of image \( I \). The image is divided into two regions \( \Omega_1 \) and \( \Omega_2 \), in region \( \Omega_1 \), \( \phi = 1 \) and \( u = c_1 \) while in region \( \Omega_2 \), \( \phi = -1 \) and \( u = c_2 \). So the piecewise constant-valued function can be defined as

\[
u = \frac{c_1}{2} (\phi + 1) - \frac{c_2}{2} (\phi - 1)
\]

where \( c_1 \) and \( c_2 \) are positive constants.

Then the energy function of the active contour model can be defined as follow:

\[
E_{\text{image}} = E_R(c_1, c_2, \phi) = \frac{1}{2} \iint_{\Omega} |u(c_1, c_2, \phi) - I|^2 \, dx \, dy + \mu \iint_{\Omega} |\nabla \phi| \, dx \, dy + \frac{1}{\tau} \iint_{\Omega} W(\phi) \, dx \, dy
\]

where \( \mu \) and \( \tau \) are the proportional coefficients. The first item is used to measure the similarity of the two-valued function \( u \) with the image \( I \), and make the function \( u \) more close to the intensity distribution of image \( I \). The second item is used to measure the length of the curve \( C \), playing the role of smoothing region boundaries. The last item is for the binary constraint.

In conventional level set methods, there is no any prior knowledge taken into account and the positive constants \( c_1, c_2 \) can be obtained directly as follow by minimizing the energy function:

\[
c_1 = \frac{\iint_{\Omega} I(1 + \phi) \, dx \, dy}{\iint_{\Omega}(1 + \phi) \, dx \, dy}, \quad c_2 = \frac{\iint_{\Omega} I(1 - \phi) \, dx \, dy}{\iint_{\Omega}(1 - \phi) \, dx \, dy}
\]

where, we can see, \( c_1 \) and \( c_2 \) are the average intensity of image \( I \) in region \( \Omega_1 \) and \( \Omega_2 \).

So when we minimize the energy function \( E_{\text{image}} \), we want the function \( u \) more close to the image \( I \), that is, the region with average intensity is close to the original image. As a result, this definition of \( u \) makes the level set model more inclined to segment out the region with consistent intensity (see Fig 2), which is similar to the thresholding segmentation method. However, the object may consist of inconsistent intensity which occurs most often in practice. Additionally, in the context of tracking, we usually have a specific target of interest, which can be explored to supervise the evolution of the curve and refine its acting orientation.

4.2. Online appearance modeling

We construct the online appearance model in a implicit manner of boosting classification as in [1]. In contrast with distinguishing between the object and background by modeling two separate models respectively, treating the separation as a binary classification problem is more inclined to seize properties of most discriminative between them, as well as decrease pollution from the similar background pixels.

The boosting-based appearance construction process can be summed up with several keys:

**Algorithm 1** On-line AdaBoost Appearance Construction

**Input:** new examples \( \{x_i, y_i\}_{i=1}^N, y_i \in \{-1, +1\} \)

**Output:** strong classifier \( H(x) \)

**Initialization:** the importance weight \( \{w_i\}_{i=1}^N = \frac{1}{N} \)

For each iteration \( t \) do:

- Make \( \{w_i\}_{i=1}^N \) a distribution.
- Choose weak classifier \( h_t(x) \) with the lowest error in the pool (for first \( K \) iterations) or train new weak classifier \( h_t(x) \) (for last \( T - K \) iterations).
- Calculate the error \( err \) and voting weight \( \alpha_t \) for the classifier \( h_t(x) \).
- Update importance weight by the classification result of classifier \( h_t(x) \).

The updated strong classifier is given by \( \text{sign}(H(x)) \), where \( H(x) = \sum_{t=1}^T \alpha_t \cdot h_t(x) \).

We apply this implicit target model as a detector in the arriving frame. For each sample evaluated, within the search region, we can obtain a confidence score indicating the likelihood it derives from the target object. In [1], mean shift algorithm is implemented for searching the peak and a rectangle is used for representing the result. However, although the tracker locks onto most competitive samples, it
may ignore the contribution of the other. Additionally, the inevitable background pixels mixed in the target region may pollute the subsequent updating results. Differently, we introduce these scores into the proposed SLSM (see Fig.3) as prior target knowledge to supervise the curve evolution and give a global optimal result of object contour, which can also supply the boosting procedure with qualified samples as well as dealing with the resale problem in mean shift algorithm.

4.3. Level set formulation

Our goal is to estimate the contour from a sequence of images. Let $I_k : \mathbf{x} \rightarrow \mathbb{R}^m$ denote the image at time $k$ 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 such that the zeroth level set of $\phi$ is $C$, that is, $\phi(x, y) = 0$ if and only if $C(s) = [x \ y]^T$ for some $s \in [0, 1]$. In response to the low efficiency of the traditional level set models, the proposed SLSM maintains the advantage of using two-valued level set function $\phi$ to replace the traditional signed distance function:

$$
\phi(x, y, k) = \begin{cases}
1, & \text{if } [x \ y]^T \text{ inside } C_k \\
-1, & \text{if } [x \ y]^T \text{ outside } C_k
\end{cases}
$$

Using this simple form can avoid the re-initialized process of the level set function in each iteration as well as the cumbersome numerical realization.

Given all the observations $I_{0:k}$ up to time $k$, score map $S_{0:k}$, and the previous contours $C_{0:k-1}$, we model the probability of the contour $C_k$ at time $k$ by considering both the region and edge cues in a Bayesian manner as

$$p(C_k | I_{0:k}, S_{0:k}, C_{0:k-1}) \propto p_{\text{region}}(S_k | C_k) p_{\text{edge}}(I_k | C_k) p(C_k | C_{0:k-1})
$$

where $p_{\text{region}}(S_k | C_k)$ presents the likelihood that the region inside and outside $C_k$ are the target object and background respectively, and $p_{\text{edge}}(I_k | C_k)$ gives the likelihood that the contour is on image edge, $p(C_k | C_{0:k-1})$ is the prior probability of the contour which we regard equally for all candidate curves. Here, the assumption we depend on is that the measurements are independent of each other.

When we maximize the probability of (9), obviously, we expect to obtain the contour that surrounds the target region and just right converges to its edge.

Let $\Omega^+$ be the region of the image inside the curve and $\Omega^-$ the region outside the curve. Let $S_k^+(\mathbf{x}) = \zeta(\mathbf{x}) |_{x \in \Omega^+}$ capture the probabilities inside $C_k$, $S_k^-(\mathbf{x}) = -\zeta(\mathbf{x}) |_{x \in \Omega^-}$ capture the probabilities outside $C_k$. $\zeta(\mathbf{x})$ denotes the score value of pixel $\mathbf{x}$ and is positive for pixels that are more likely to belong to the target than to the background, and vice versa for negative. The region-based probability $p_{\text{region}}(S_k | C_k)$ can be decomposed as

$$p_{\text{region}}(S_k | C_k) \propto p_t(S_k^+ | C_k) p_b(S_k^- | C_k)
$$

where

$$\log(p(S_k^+ | C_k)) \propto \sum_{x \in \Omega^+} T(x, y)
$$

$\log(p(S_k^- | C_k)) \propto \sum_{x \in \Omega^-} T(x, y)$

where $* \text{ denotes } +,- \text{ for } p_t \text{ and } p_b \text{ respectively}$. Under the objective of driving the contour to the target boundary, we use image gradient for edge detecting, see Fig.3, and the edge-based probability $p_{\text{edge}}(I_k | C_k)$ can be computed as

$$p_{\text{edge}}(I_k | C_k) \propto \sum_{[x \ y]^T \in C_k} T(x, y)
$$

$T(x, y) = |\nabla[G_\sigma(x, y) * I_k(x, y)]|^2$

where $\nabla$ denotes spatial gradient operator, * denotes convolution and $G_\sigma$ is the Gaussian filter with standard deviation $\sigma$.

We define the energy function, minimizing which over the level set function is equivalent to maximizing the probability of (9), as follow:

$$E(\phi) = \int_{\Omega^+} -S^+(\mathbf{x})d\mathbf{x} + \int_{\Omega^-} -S^-(\mathbf{x})d\mathbf{x} + \xi \int_{\Omega} -T(x)d\mathbf{x}
+ \mu \ell(C) + \frac{1}{\tau} \int_{\Omega} W(\phi)d\mathbf{x}
$$

where $\xi, \mu \text{ and } \tau$ are the coefficients that weight the relative importance of each item. $\ell(C)$ is the length of the curve.

The last item is for constraint of $\phi^2 = 1$, where $W$ can be defined as $(\phi^2 - 1)^2$ and $\Omega = \Omega^+ \cup \Omega^-$ is the image domain.

Employing the binary level set function as a differentiable threshold operator, we unify the integral region and rewrite (14) as

$$E(\phi) = \int_{\Omega} -S^+(\mathbf{x})(1 + \phi) - S^-(\mathbf{x})(1 - \phi) - \xi T(\mathbf{x})(1 - \phi^2)
+ \mu |\nabla \phi| + \frac{1}{\tau} \int_{\Omega} W(\phi)d\mathbf{x}
$$

where $\ell(C) = \int_{\Omega} |\nabla \phi|d\mathbf{x}$. The associated Euler-Lagrange equation for this function can be given by

$$0 = -S^+(\mathbf{x}) + S^-(\mathbf{x}) + 2\xi T(\mathbf{x})\phi - \mu \text{div} \left(\frac{\nabla \phi}{|\nabla \phi|}\right) + \frac{1}{\tau} \int_{\Omega} W'(\phi)d\mathbf{x}
$$
and implemented by the following gradient descent:

$$\frac{\partial \phi}{\partial t} = S^+(x) - S^-(x) - 2\zeta T(x)\phi + \mu \text{div}(\frac{\nabla \phi}{|\nabla \phi|}) - \frac{1}{\tau} W'(\phi)$$

(17)

where $\text{div}$ is the divergence operator.

In contrast with conventional level set formulation, ours, instead of based upon intensity edges, is supervised by the specific knowledge of the target. Therefore, the curve, in SLMS, can be steered to the target from a wide variety of states, without any request of the initial curve that must be inside or outside the target completely. Fig.3 illustrates the tracking principle of the proposed algorithm.

5. Experimental results

In this section, the proposed method was tested on several video sequences with different challenges for tracking. All the sequences were captured by moving cameras. In all cases, we use five weak classifiers and employ the local histogram of oriented gradients as well as the color cues to construct the 11D feature vector as in [1]. 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.

The first sequence consists of 820 frames and describes a ship, with similar color distribution to the water, navigating on the river with moving waves behind and illumination changes. We can see the tracking performance of our method is good as shown in Fig.4.

In above test, many contour tracking approaches can give good results as the proposed method and the same phenomenon can be observed on other unicolor target sequences. Then, further more, we compared the proposed method with some other algorithms on a multi-mode target sequence to show the improvement of our approach. This sequence describes a woman with multi-colored appearance walking in a clutter street with large posture changes and frequent sheltering cases. The three algorithms we tested are: (a) standard particle filter which uses HSV color histogram as the appearance model in [16]; (b) traditional Mumford-Shah method within the particle filter framework for contour tracking in [18]; (c) the proposed method. Fig.5 shows the tracking results of these algorithms. We can see that in standard particle filter, the candidate states are roughly generated and do not guarantee to pick up the whole target accurately. The algorithm (b), based on conventional level set model, considers only intensity edges without any target information which results in its convergence in multi-colored region highly depending on the initial curve. Therefore, incompetent results are shown on multi-colored object
tracking due to the unreliable initial curves derived from the prediction step of particle filter framework. The proposed algorithm, in contrast, improves the tracking quality dramatically due to the supervising of the specific target knowledge.

Unlike the explicit target model methods, the classifier based implicit representation has the ability of accommodating information of target in the past as well as dealing with occlusion cases. We use decrease rate of the detected positive samples as an occlusion detector. Once detected, we slow down the speed of updating the weak classifiers as in [1], allowing tracking to resume when the target reappears as shown in Fig.5(c).

Another three challenging sequences were tested to further evaluate the proposed method. The first described a man in strip colorful clothes walking on the balcony, undergoing significant scale changes and shape deformation as he walked toward or deviating from the camera. It is a challenge for traditional intensity edge based level set methods to represent the person accurately. As we can see in Fig.6(a), the proposed method shows pleasant tracking results, demonstrating the effectiveness of the technical. The second sequence contained a toy lemming moving fast above the table with a clutter background behind as well as sheltering cases. From the tracking results shown in Fig.6(b), we can see that our method performs well even in complicate scene. The third sequence described a bottle of liquor taken and circled around another similar one. It is a challenge for the method of [5] where the target and background are segmented based on intensity consistent fragments and separately modeled in a GMM manner. In contrast, our method is classifier based and highlights the discrimination between the object and its local neighbor. In this test, we use the patch based classifier [8] so as to seize the global characteristics of the target. Fig.6(c) shows the tracking results of the third sequence, indicating the robustness of the proposed method in dealing with these challenging cases.
As we have demonstrated, the tracker, so far, can manage to overcome partial occlusions automatically. However, the object may undergo a long period of complete occlusion in real-world applications. We use a simple mechanism to handle these cases. Once the number of detected positive samples drops sharply and the object is determined to be occluded, we stop updating the target model. As long as the target is not visible, the particle filter procedure is used to predict the object position while the resulting contour of the second frame is used to hallucinate the object shape. Once the particle filter finds the positive region when the target reappears, tracking resumes. Two video sequences involving occlusion were tested. These sequences describe a toy tiger and star patch which were pulled across the table and occluded by the mouse pad for a long period. Fig. 7 shows the tracking results of the proposed method on these two sequences, from which we can see the proposed method is competent to deal with occlusion cases. Note that, we use the resulting contour of the second frame to hallucinate the invisible object shape here for the consideration of computational simplicity which is also acceptable for these cases due to the little scale changes of the targets. One can further improve it by using the most similar contour over the previous output with the one just prior to the occlusion.

Figure 6. Experimental results of further evaluating.

Figure 7. Experimental results involving complete occlusion.
Finally, we show the ability of the proposed method to work on gray scale images, which are usually difficult to track because a single color channel does not provide enough information for tracking. Fig. 8 shows the tracking results of a gray scale video sequence. This sequence describes a pop can which is held and swayed round a bunch of plant. As the pop can moves and turns, appearance and illumination changes as well as severe occlusion occur. 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 supervised level set model (SLSM) in this paper for non-rigid object tracking. By considering the context of tracking, we refined the curve evolution by the specific knowledge of the target we want to track, which is learnt in an online boosting manner. In contrast with conventional intensity edge based level set methods, our approach is object-oriented and could lead an accurate convergence to the targets in tracking applications. Experimental results on several challenging video sequences have verified that the proposed method is effective in many complicate scenes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61071180) and Space Science Foundation (No. 20105577015).

References