KERNEL BASED ARTICULATED OBJECT TRACKING WITH SCALE ADAPTATION AND MODEL UPDATE

Anbang Yao[a], Guijin Wang[a], Xinggang Lin[a], Hao Wang[b]

[a]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
[b]Nokia Research Center, Beijing, 100013, China

ABSTRACT

Kernel based object tracking (KBOT) is one of the most popular and effective techniques for tracking task. However the constancy of the target model and unsound scale adaptation method are two main limitations. In this paper, we present a kernel based approach incorporated with scale estimation and target model update for articulated object tracking task. After predicating the object center with scale fixed KBOT, we extend scale selection theory to estimate the local optimal object scale. Once the object scale has been estimated, a kernel density estimation based strategy is developed to update the target model. Experimental results show that our approach is superior to traditional KBOT in the following two aspects: 1) it is less affected by the object scale change; 2) it is less prone to appearance variation.

Index Terms—Articulated object, Kernel based object tracking, Kernel density estimation, Scale space

1. INTRODUCTION

Video-based object tracking is of great pertinence to today's emerging applications such as visual surveillance [1], intelligent traffic navigation [2], human computer interaction [3], video indexing [4] and content based video compression [5] etc. Among numerous approaches [6-8], kernel based object tracking (KBOT) [9] has recently gained more and more attention among tracking community due to its low computation cost, easy implementation and competitive performance.

Since articulated objects [9-11] always appear in a wide range of appearances and scales, KBOT has two main limitations: the constancy of the target model and unsound scale adaptation scheme. As for scale change, KBOT [9] first assumes that the object scale keeps in constancy during the course of tracking. Based on the largest Bhattacharyya coefficient, a three scales select one strategy (they used ±10%) is offered, which usually makes the object scale grow too large or shrink too small. Yilmaz's work [8] relies on the asymmetric kernel based mean shift to estimate object scale and orientation. However, it is not trivial to generate appropriate kernel for articulated objects. In addition, only relatively rigid object is concerned in [8]. Collins [12] extends the KBOT by combining a parallel procedure to compute the bandwidth of the kernel in the scale space. This approach demands both object center and kernel bandwidth for convergence in an interleaving mode, which is difficult in some scale clutter cases [13]. As for appearance variation, target model should be updated properly with a predefined similarity measure. The popular similarity measures are the Bhattacharyya coefficient [9] and the Kullback-leibler divergence [7]. However，both measures must resort to the discriminative features and are not suitable for high dimensional cases [13].

In this paper, we present a robust kernel based articulated object tracking approach incorporated with scale adaptation and target model update. The contributions of our scheme are two aspects: 1) we extend scale space theory [14] in a different way to adapt to changing object scale; 2) we develop kernel density estimation [15] as a measure to adaptively update target model.

The remainder of this paper is organized as follows. Section 2 presents a detailed description of our approach, including the motivation of scale fixed KBOT, an insight into our scale space based approach for estimating object scale and target model update strategy. Complete implementation of our approach is presented in Section 3. The effectiveness of the proposed approach is demonstrated in Section 4. Section 5 concludes with a brief remark.

2 PROPOSED TRACKING APPROACH

In visual tracking, a tracker is tailored for assigning consistent labels to the tracked object in each frame of a video [3]. Three main modules of our tracker will be described below in consecutive order.

2.1. Scale fixed KBOT

This work is jointly supported by Nokia Research Center, Beijing, China and National Natural Science Foundation of China under Grant No. 60472028 and Doctoral Program Research Foundation of Ministry of Education of China under Grant No.20040003015.

In our scheme, we first obtain the object center of each frame using scale fixed KBOT. The basic concept of scale fixed KBOT is a simpler version of [9]. Given a present target model q_u and a candidate region model p_u at 2D center x_0 in the subsequent frame, the aim of KBOT is to consistently estimate a new center $x^{'}$ in the subsequent frame. To this end, the similarity objective function in KBOT to be maximized is

$$\rho(p(x),q)=\sum_{u=1}^{M}\sqrt{p_u(x)q_u}\ . \tag{1}$$

Using the Tailor expansion at values $p_u(x_0)$, the linear approximation of Bhattacharyya coefficient (1) is obtained as

$$\rho(p(x),q)\approx\frac{1}{2}\sum_{u=1}^{M}\sqrt{p_u(x_0)q_u}+\frac{\sum_{i=1}^{N}w_i k(\left\|\frac{x-x_i}{h}\right\|^2)}{2\sum_{i=1}^{N}k(\left\|\frac{x-x_i}{h}\right\|^2)}, \tag{2}$$

where $k(\cdot)$ is a non-negative, non-increasing and piecewise differentiable kernel profile, M is model dimension, N represents the number of pixels (each pixel value is donated by $b(x_i)$) located in $k(\cdot)$, h is 2D bandwidth of $k(\cdot)$, and

$$w_i=\sum_{u=1}^{M}\sqrt{\frac{q_u}{p_u(x_0)}}\delta\left(b(x_i),u\right). \tag{3}$$

To maximize (2), its derivative with respect to x is computed firstly. Setting resulted derivative equal to zero, mean shift based optimization is utilized and the iterative expression is deduced as follows

$$x^{'}=\frac{\sum_{i=1}^{N}x_i w_i g(\left\|\frac{x_i-x_0}{h}\right\|^2)}{\sum_{i=1}^{N}w_i g(\left\|\frac{x_i-x_0}{h}\right\|^2)}, \tag{4}$$

where $g(\cdot)$ is the derivative of $k(\cdot)$. In our approach, elliptical profile function is used which makes $g(\cdot)$ is a constant, and the bandwidths of $g(\cdot)$ are fixed at each iteration. As for models, RGB based position weighted histogram is adopted.

2.2. Scale selection

Usually, scale fixed KBOT performs well on object position tracking task, just like Collins' work [12] demonstrated. But it does nothing to adapt to the object scale variation, which will further worsen the estimated object center during tracking period. To modify these disadvantages, one possible approach is to search for object scale by evaluating the distances over a set of different scales. But this method usually makes object scale shrink too small or grow too large [12]. Another way is to implement a brute-force search which is not practical due to its high computational cost. Therefore, how to perform the accurate search with considerable computation complexity is of great importance. Alternatively, the mean shift algorithm can also be carried out to estimate a local optimal kernel scale in a suitable space which corresponds to kernel functions.

Hinted at the promising properties of scale space theory [14], we adopt scale space as the corresponding space to construct the mean shift based iterative procedure. There are a number of reasons for choosing this space. First, scale space theory provides a robust and effective way to select the best scales for discriminating features in image plane. Second, Gaussian kernel is the unique kernel to generate a scale space, which can be computed in particularly efficient mode (e.g. a pre-calculated look-up-table). In addition, Gaussian function is more stable compared with other functions, e.g. sigmoid function.

To visualize the aimed iterations, we also need to define a scale distribution s (e.g. $s_n=1\pm x^n$, where $0<x<1$, and $n\in\mathbb{N}$). Treating the scale kernel function D with respect to s, we find that

$$D(x,s)=G((x-x_0)/h,s\sigma_0/\sqrt{1.6})-G((x-x_0)/h,\sqrt{1.6}s\sigma_0)$$
$$where\quad G(x,\sigma)=\frac{1}{2\pi\sigma^2}e^{-x^2/2\sigma^2} \tag{5}$$

Since w is computed by position weighted histograms which combine color values and spatial positions in a unified way so as to avoid the need for modeling object shape, appearance or motion, we also take w as weight vector to estimate scale. Now the mean shift based iterations can be performed by (6) until it converges to a local optimal solution.

$$s^{'}=\frac{\sum_s\sum_{x_i}\left|D(x_i,s)\right|w(x_i)s}{\sum_s\sum_{x_i}\left|D(x_i,s)\right|w(x_i)}. \tag{6}$$

Note that, the updating rule (6) is quite similar to (4), it can be easily converged in a fast way. Another merit we want to emphasize is that (6) also strongly bears the properties of scale space theory for elegant scale selection.

2.3. Target model update

To achieve consistent tracking performance, appearance variation of the articulated object must also be concerned. As mentioned above, the typically used similarity measures are the Bhattacharyya coefficient [9] and the Kullback-leibler divergence [7]. But in practice, these similarity measures usually face two difficulties. First, both measures must resort to the discriminative features. However, the common used features like color histogram and motion cue lose the spatial information of the object, which highly degrades the measures' effectiveness. Second, the classical similarity measures are not discriminative for high dimensional cases [13]. To deal with these difficulties, kernel density estimation technique is utilized as a similarity measure for updating target model in our method.

Denoting an original target model as $a=\{x_i^a, u_i^a\}_{i=1}^N$ and a new tracked target region as $b=\{x_j^b, u_j^b\}_{j=1}^M$, where a, b represent the sets of sample points in each region, each sample point is represented by a 2D position vector x and a feature vector u (e.g. three color components). It should be emphasized that this kind of sample point combines spatial information and traditional features together and provides better discrimination ability. To measure the similarity between two regularized regions, the similarity function is defined as

$$p(a,b)=\frac{1}{NM}\sum_{j=1}^{M}\sum_{i=1}^{N}W(\frac{x_j^b-x_i^a}{\sigma})K(\frac{u_j^b-u_i^a}{h}), \quad (7)$$

where $W(\cdot)$ is the position kernel function with a bandwidth σ, while $K(\cdot)$ is the feature kernel function with a bandwidth h. The position kernel function $W(\cdot)$ should satisfy $W(x)>0$ and $\int_{-\infty}^{+\infty}W(x)=1$, and the same to $K(\cdot)$. As for kernel function selection, Gaussian function is usually used, and one look-up-table can be calculated offline to highly save computation cost. According to (7), the distance measure is expressed as

$$d(a,b)=\sqrt{1-p(a,b)}. \quad (8)$$

With distance measure (8), the updating process is implemented in a straightforward and simple form. That is

$$\begin{cases} select\ a & if\ d(a,b)<T \\ select\ b & others \end{cases}. \quad (9)$$

In our approach, both the original target model and the new tracked target region in the current frame are represented by an elliptical region. Coordinated positions and corresponding RGB color values are used in a unified way. To eliminate the influence of the different sizes of two regions, the new tracked target region is normalized to the same size as the original target model. Taking these two normalized regions as possible target model candidates, we select the more appropriate one as target model for visualizing the tracking task in the next frame by (9).

3 IMPLEMENTATION OF ALGORITHM

To summarize, we outline the proposed algorithm as below.

1. Initialize target model q [9], $s=\{0.85, 0.95, 1, 1.05, 1.15\}$.
2. Tracking with fixed scale:
   a) Keep object scale fixed, donate x_0 as object center in frame n, generate target candidate model p [9] at x_0.
   b) Compute weight vector w by (3).
   c) Perform mean shift based iterative procedure using (10). Stop iteration until object center is converged.

$$x'=\frac{\sum_{x_i}w(x_i)x_i}{\sum_{x_i}w(x_i)}. \quad (10)$$

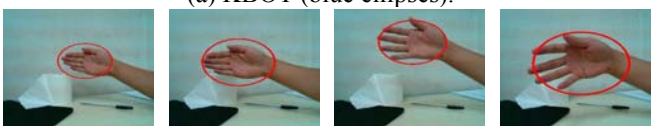
3. Select the local optimal object scale by iterating (6).
4. Update target model q by (9).
5. Set $x_0=x'$, $s=s's$, $n=n+1$, go to step 2.

4 EXPERIMENTS

To demonstrate the performance of our approach, two hand sequences are captured at 15 frames per second by a web camera (worse quality with blurring) in lab environment with the image resolution of 320×240 pixels. We compare our approach with KBOT offered by [9].

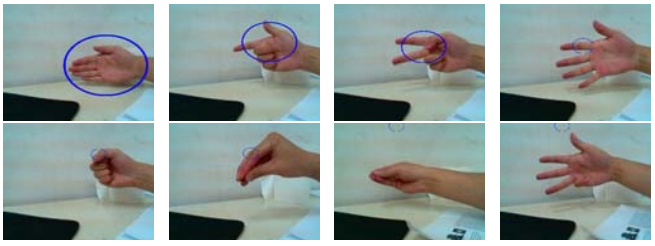
The shorter sequence 1 has 285 frames, in which the hand mainly undergoes scaling. Figure 1 shows some result examples. Without model update module, our approach can accurately reach hand center and correctly adjust the kernel scale throughout the entire sequence. While for KBOT, the resulted ellipse begins to drift away from hand center at frame 35 and to heavily shrink at frame 127.

(a) KBOT (blue ellipses).

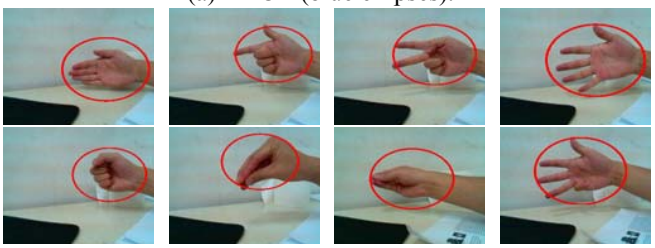


(b) Our approach (red ellipses).

Fig.1 Tracking result examples of sequence 1. From left to right, the frame indexes are 1, 35, 127 and 219.



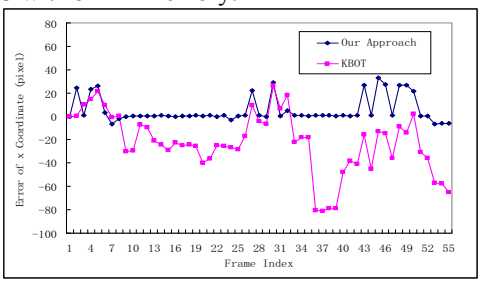
(a) KBOT (blue ellipses).



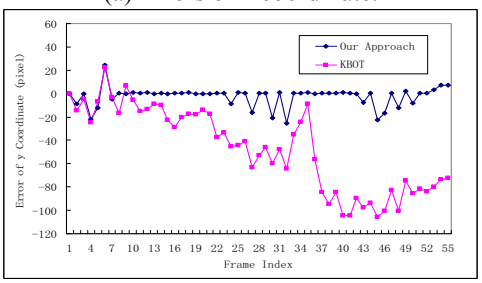
(b)Our approach (red ellipses).

Fig.2 Tracking result examples of sequence 2. From left to right, the frame indexes are 1, 293, 357, 457, 663, 1017, 1185 and 1525.

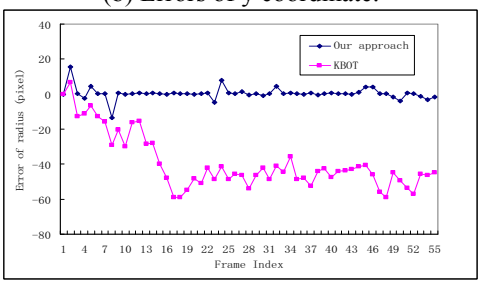
The long sequence 2 of 1636 frames is even more difficult, in which the hand undergoes scale changing, more than 10 different postures, and translation with some in-plane rotation. Some result examples are given in Figure 2. Note that KBOT fails quickly when the hand appearance is changed. Adopting the proposed model update strategy, our approach can deal with all these difficulties mentioned above in a robust and precise way. Figure 3 gives tracking errors of x coordinate, y coordinate and object scale. It can be seen that our approach yields stable center positions and accurate scales. The small displacements of our curves are mainly due to the bias resulted from the difficulty of labeling ground truth. In addition, our approach achieves up to a processing speed of 40 fps in average on a Pentium 2.5 GHZ PC with 512M memory.



(a) Errors of x coordinate.



(b) Errors of y coordinate.



(c) Errors of the hand radius (short axis).

Fig. 3 Errors of each frame in sequence 2. Each frame is selected in every 30th frame.

5 CONCLUSIONS

To persistently track an articulated object with changing appearance and scale, the target model should adapt to appearance variation and the object scale must also be adaptive. Since scale fixed KBOT performs well in tracking an object with less varying appearance, we construct a scale space based method to estimate the object scale after the object center is obtained. By combining spatial and feature cues in a unified way, a kernel density estimation based strategy is developed to update the target model. Experimental results verify that our approach completely outperforms KBOT in scale and appearance changing scenarios, and real time computation cost is also achieved.

6. REFERENCES

[1] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9), pp. 1208-1221, 2004

[2] W.M. Hu, X.J. Xiao, D. Xie, T.N. Tan and S. Maybank, "Traffic accident prediction using 3-D model-based vehicle tracking", *IEEE Transactions on Vehicular Technology*, 53(3), pp. 677-694, 2004.

[3] B. Stenger, A. Thayananthan, P.H.S. Torr and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9), pp. 1372-1384，2006

[4] Z.Z. Yin and R. Collins, "On-the-fly object modeling and tracking", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[5] A. Eleftheriadis and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconference sequences at low bit rates", *Signal Processing: Image Communication*, 7(3), pp. 231-248, 1995.

[6] A. Yilmaz and M. Shah, "Object tracking: a survey", *ACM Computing Surveys*, 38(4): 45 pages, 2006.

[7] A. Yilmaz, "Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, pp. 728-731, 2007.

[8] M. Isard and A. Blake, "CONDENSATION—conditional density propagation for visual tracking", *International Journal of Computer Vision*, 29(1):5-28, 1998.

[9] D. Comaniciu, V. Ramesh and P. Meer, "Real-time tracking of non-rigid objects using mean shift", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 142-149, 2000.

[10] Z.M. Fan, Y. Wu and M. Yang, "Multiple collaborative kernel tracking", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 502-509, 2005.

[11] O. Masoud and N.P. Papanikolopoulos, "A novel method for tracking and counting pedestrians in real-time using a single camera", *IEEE Transactions on Vehicular Technology*, 50(5), pp. 1267-1278, 2005.

[12] R.T. Collins, "Mean-shift blob tracking through scale space", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 234-240, 2003.

[13] C.J. Yang, R. Duraiswami and L. Davis, "Efficient mean-shift tracking via a new similarity measure", *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 176-183, 2005.

[14] T. Lindeberg, "Feature Detection with Automatic Scale Selection", *International Journal of Computer Vision*, 30(2), pp. 79-116, 1998.

[15] A. Elgammal, R. Duraiswami, D. Harwood and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance", *Proceedings of the IEEE*, vol. 90, pp. 1151-1163, 2002.