IEICE TRANS. INF. & SYST., VOL.E94–D, NO.4 APRIL 2011

LETTER

Non-rigid Object Tracking as Salient Region Segmentation and Association

Xiaolin ZHAO[†a)], ***Student Member***, **Xin YU**[†b)], **Liguo SUN**[†c)], **Kangqiao HU**[†d)], ***Nonmembers***, **Guijin WANG**[†e)], ***Member***, ***and*** **Li ZHANG**[†f)], ***Nonmember***

SUMMARY Tracking a non-rigid object in a video in the presence of background clutter and partial occlusion is challenging. We propose a non-rigid object-tracking paradigm by repeatedly detecting and associating saliency regions. Saliency region segmentation is operated in each frame. The segmentation results provide rich spatial support for tracking and make the reliable tracking of non-rigid object without drifting possible. The precise object region is obtained simultaneously by associating the saliency region using two independent observers. Our formulation is quite general and other salient-region segmentation algorithms also can be used. Experimental results have shown that such a paradigm can effectively handle tracking problems of objects with rapid movement, rotation and partial occlusion.

key words: *non-rigid object tracking, saliency region segmentation*

1. Introduction

Non-rigid object tracking is still a challenging problem in computer vision. It is also an important issue in animation, behavior analysis, visual surveillance and so on. The challenges of non-rigid object tracking generally arise from tracking drift, which is usually caused by object shape variation and partial occlusion.

To handle the challenges, a variety of algorithms have been proposed. Mean-shift [1] is a powerful, non-parametric statistical method for non-rigid object tracking. The shape of a tracked object is often approximated by an ellipse or rectangle. Although being successful in some applications, it suffers from the inability to properly adapt the ellipse/rectangle when the shape and size of the tracked object change [2]. An inaccurate ellipse/rectangle often contains background area which can lead to tracking drift.

In order to make the tracker more robust to drift, high-level knowledge was utilized to model the object [3]–[5]. In [3], [4], discriminative local shape features selected by a boosting algorithm are used to detect object. To some degree, the above approaches have achieved promising performance in tracking. But these algorithms still simply approximate the object with rectangle or ellipse. Such a simple approximation has trouble tracking non-rigid objects in cluttered scenes.

In recent studies [6]–[8], researchers confirmed that using visual saliency can substantially improve segmentation and tracking performance. Saliency based detection/segmentation framework achieves promising results, especially for objects that can't be well approximated by an ellipse or rectangle. In Donoser's approach [6], local MSER (Maximally Stable Extremal Region) detector is combined with data association to track objects. Fukuchi [7] introduced MAP-based framewise segmentation based on the maximum posterior estimation of the Markov random field and visual saliency. Although the results of above approaches are promising, they may have trouble when the background is clutter and occlusion happens.

In order to track non-rigid object in clutter background, we proposed a novel strategy for non-rigid object tracking as shown in Figs. 1 and 2. Segmentation is operated in each frame by detecting saliency regions utilizing basic image cues (gray value). The precise object tracking result is obtained by finding the best corresponding regions between the current and previous frame while using two independent observers.

The rest of this paper is organized as follows: Sect. 2 introduces our method. Experimental results are shown in Sect. 3 and conclusions are drawn in Sect. 4.

2. Tracking as Salient Region Segmentation and Association

Tracking can be considered as an inference problem to locate a specific object in each frame. Typical tracking approaches aim at finding the location of the object center. We endeavor to obtain not only the location of an object's center but also the accurate object mask. Salient region segmentation is operated independently in each frame using low-level image features. The high-level knowledge of object is obtained from object data set trained by a Gentle Adaboost. Once salient region segmentation results are obtained, the low-level and high-level cues are combined together to find the best candidates as object region.

2.1 Salient Region Segmentation

In general, any of the region-of-interest (saliency or ROI) segmentation or detection methods can be used. We adopt

Manuscript received October 1, 2010.
Manuscript revised December 19, 2010.

[†]The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
a) E-mail: zhaoxiaolin00@mails.tsinghua.edu.cn
b) E-mail: xin-yu09@mails.tsinghua.edu.cn
c) E-mail: slg00@mails.tsinghua.edu.cn
d) E-mail: hkqdtc1@yahoo.cn
e) E-mail: wangguijin@tsinghua.edu.cn
f) E-mail: chinazhangli@tsinghua.edu.cn

DOI: 10.1587/transinf.E94.D.934

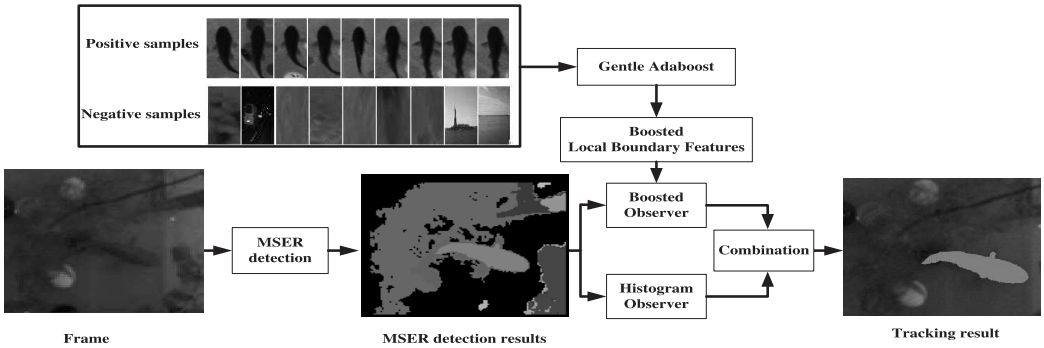

Fig. 1 Overview diagram of our approach.

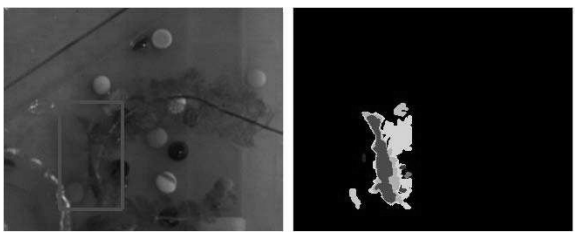

Fig. 2 (a) boundary box initialization (b) saliency region segmentation results (Only the results in boundary box area are shown). Different salient regions are filled with different color.

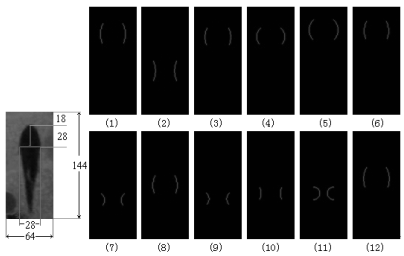

Fig. 3 The top 12 edgelet features selected by Gentle Adaboost. (The edgelets are corresponding to fish head and torso)

MSERs detection method proposed by Matas et al. [9]. MSERs has two important properties: it is invariant to affine intensity changes and can be detected at different scales. Given an image $I(x)$ (x is the pixel position), we denote the set of all extremal regions in image I as $\mathcal{R}(I)$. As in [9], the maximum pixel value in every extremal region R can be defined as:

$$I(R) = \max_{x \in R} I(x) \tag{1}$$

Define $\triangle > 0$, $R_{-\triangle}$ is the biggest extremal region is contained in R with intensity lower than R by at least $\triangle$.

$$R_{-\triangle} = \max_{U \in \mathcal{R}(I), U \subset R, I(U) \leq I(R) - \triangle} |U| \tag{2}$$

Similarly, define $R_{+\triangle}$ as following:

$$R_{+\triangle} = \min_{U \in \mathcal{R}(I), U \supset R, I(U) \geq I(R) + \triangle} |U| \tag{3}$$

$R_{+\triangle}$ is the smallest extremal region containing R and has intensity which exceeds at least $\triangle$ of R. $|.|$ is the size of the region. Denote the area variation as:

$$v(R, \triangle) = \frac{|R_{+\triangle}| - |R_{-\triangle}|}{|R|} \tag{4}$$

The region R is a MSER when Eq. (4) takes a minimum.

2.2 Boosting Discriminative Boundary Fragments

Discriminative boundary fragments which can capture the local shape of an object are selected by Gentle Adaboost. A discriminative boundary fragment (also named edgelet in [3]) is a short segment of line or curve. Define the positions and normal vectors of the elements in an edgelet as $\{u_i\}_{i=1}^{l}$ and $\{n_i^E\}_{i=1}^{l}$, where l is the length of the edgelet. The affinity between the edgelet and the image I at position w can be calculated by

$$S(w) = (1/l) \sum_{i=1}^{l} M^I(u_i + w) | < n^I(u_i + w), n_i^E > | \tag{5}$$

The edge intensity $M^I(p)$ and normal vector $n^I(p)$ of I are calculated by 3×3 Sobel kernel convolution. The selection results are shown in Fig. 3.

2.3 Combining Low-Level and High-Level Cues

We adopt a generative gray histogram observer model and a boosted observer model to describe the object. Both observers work on gray scale data. The proposed approach embeds the low-level and high-level knowledge of the object into the observation models.

2.3.1 A Histogram Observer (Low-Level Cue)

R_o^{t-1} denotes the object region in previous frame and q denotes the corresponding object histogram. With p denoting the histogram of an extremal region R^t in frame I_t, the Bhattacharyya distance between the two regions can be defined as

$$d(R_o^{t-1}, R^t) = \sqrt{(1 - \rho[p, q])} \tag{6}$$

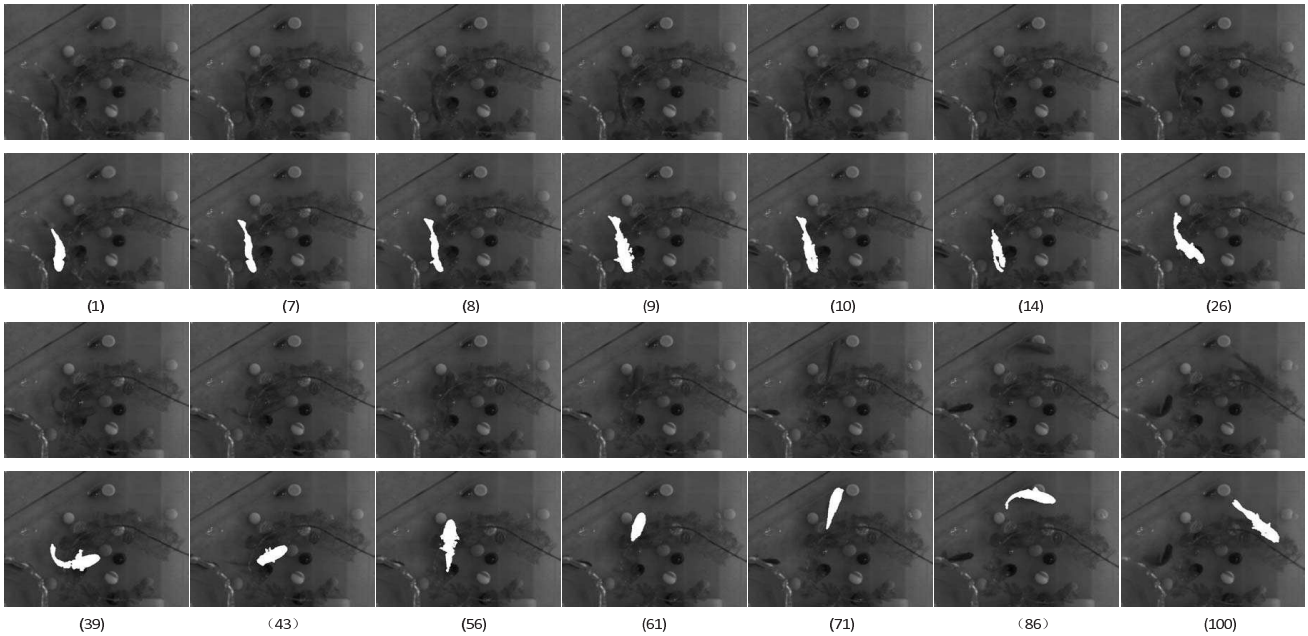

Fig. 4 Tracking results with our approach. The first and third rows are the original video frames, the other rows are the corresponding tracking object regions.

where we chose $\rho[p,q] = \sum_{b=1}^{B} \sqrt{(p_b q_b)}$. B is the number of histogram bins, and b is the histogram bin. The observation likelihood can be modeled by a Gaussian distribution:

$$p(R^t|R_o^{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-d(R_o^{t-1}, R^t)^2}{2\sigma^2}\right) \quad (7)$$

where σ^2 is the variance of Gaussian distribution.

2.3.2 A Boosted Observer (High-Level Cue)

This observer is boosted from a pool of Linear Discriminant Analysis (LDA) classifiers. The observation likelihood is modeled by a Sigmoid function of the boosted output:

$$p(R|M_{edgelets}) \propto \frac{1}{1 + exp\left(-\frac{\sum_k \alpha_k sign(w_k^T f_k(M_{edgelets}) - \eta_k)}{\sum_k \alpha_k}\right)} \quad (8)$$

where $(\alpha_k, w_k, f_k, \eta_k)$ is the $k-th$ weak classifier. f_k and α_k are the features and the corresponding boosted weight, w_k and η_k are the LDA projection vector and threshold. $M_{edgelets}$ are the top 12 edgelet features selected by Adaboost.

2.4 Associating by Maximum A Posteriori (MAP) Estimating

The most likely object region $\widehat{R^t}$ is obtained by MAP estimation with an assumption that the observers are independent:

$$\widehat{R^t_{ML}} = \underset{R^t \in \mathcal{R}(I_t)}{argmax}\ p(R^t|R_o^{t-1}) p(R^t|M_{edgelets}) \quad (9)$$

3. Experimental Results

Basically, fish tracking epitomizes the most problems in non-rigid object tracking, such as shape change, less textures, rotation and occlusion. So we chose fish tracking to evaluate our tracking algorithm.

3.1 Implementation Details

The salient regions extracted in Hue channel are normalized before observing using method proposed in [10]. The grey histogram model is updated online. Its likelihood is evaluated by a Gaussian model. Histogram bin number $B = 16$ and the variance $\sigma = 1$. The length of one edgelet is from 4 pixels to 16 pixels. Because fish shape is symmetry so we focus on the 1/8 circles, 1/4 circles and their symmetric pairs. The windows size is 64×144 pixels and the overall number of edgelets is 1,396,480.

In this experiment, we do not use any scene structure or background subtraction to facilitate segmentation. The image size of the test sequence is 640×512 pixels. Our experimental machine is a dual-core dual-processor Intel Pentium 2.4 GHz CPU. The presented results were implemented in C++, enabling a tracking speed of about 35 frames per second for the test sequence. The initial region is detected by MSER detection algorithm in a rectangle area which is selected by hand in first frame. In all experiments the MSER detection parameters $\triangle$ were fixed to 1 and the bounding box size of segmentation area is double of the object size.

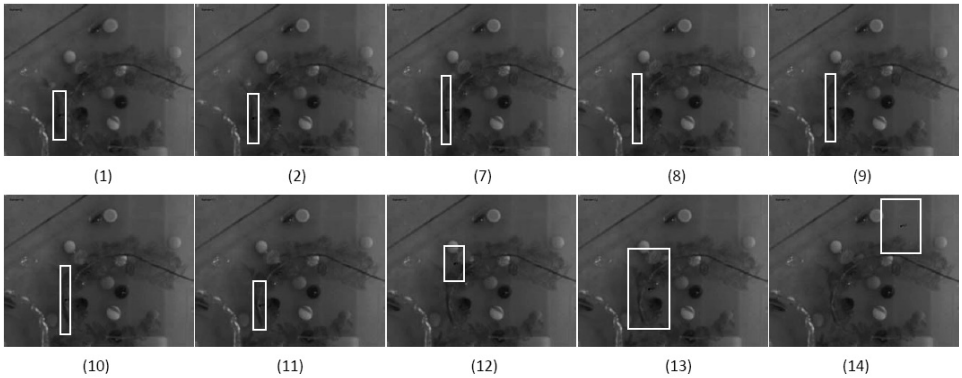

Fig. 5 Tracking results with MSER Tracker [6]. It drifted to background in the 12th frame.

3.2 Results

Our goal is to make our algorithm work well on non-rigid object tracking. Among 46 fish in 18 video sequences (over 16,000) frames, after assigning the initial positions, our tracking algorithm successfully tracks 39 fish, achieving a correction rate ($r = N_{right}/N_{total}$, N_{right} is the number of correct tracked frames and N_{total} is the number of total frames) of 84.7%. The correction rate for MSER tracker [6] is 78.8%.

Figure 4 is an example of the tracked fish results. There are some small stones and weeds in the background. A fish moved in the scenario. Frequently partial occlusion further increased the difficulty of tracking. The partial occlusion makes the salient segmentation can not provide an integral shape for the observers. Although the two observers can only observe parts of the object in the 7-th, 8-th, 43-th, 61-th frames, our tracker still found the reasonable salient regions as the tracking results. Utilizing low-level image information ability enables the tracker re-obtain the whole object region when the partial occlusion disappears. From 71-th frame we can find that the object integral shape was re-segmented out after occlusion disappeared. In addition, we compare our results with the [6] proposed tracker. Although it succussed in some applications, it encounters with drift problem when background has similar regions for the usage of only the low-level features of object which can be seen in Fig. 5.

4. Conclusion

In this paper, we proposed a novel non-rigid object tracking paradigm, which combines salient region segmentation with discriminative observers. Experimental results on fish sequence have validated its robustness to objects with shape variation, partial occlusion and background clutter. Although our experiments are only about fish tracking, this paradigm can also be extended to other non-rigid object tracking.

Acknowledgement

The authors would like to thank anonymous reviewers for constructive comments that help improve the quality of this manuscript.

References

[1] D. Comaniciu and V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," Proc. IEEE Conf. Comp. Vision Pattern Recognition, vol.2, pp.142–149, 2000.
[2] H. Zhou, T. Liu, J. Zheng, et al., "Tracking non-rigid objects in video sequences," Int. J. Inform. Acquisition, vol.3, no.2, pp.131–137, 2006.
[3] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," Proc. IEEE Int. Conf. Comp. Vision, vol.1-2, pp.90–97, 2005.
[4] A. Opelt, A. Pinz, and A. Zisserman, "A boundary-fragment-model for object detection," Proc. European Conf. Computer Vision, vol.2, pp.575–588, 2006.
[5] F. Jiang, G.J. Wang, C. Liu, et al., "Robust object tracking via combining observation models," IEICE Trans. Inf. & Syst., vol.E92-D, no.1, pp.1–4, Jan. 2009.
[6] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (MSER) tracking," Proc. IEEE Conf. Comp. Vision Pattern Recognition, vol.1, pp.553–560, 2006.
[7] K. Fukuchi, K. Miyazato, A. Kimura, et al., "Saliency-based video segmentation with graph cuts and sequentially updated priors," Proc. IEEE Int. Conf. on Multimedia and Expo., pp.638–641, 2009.
[8] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," Proc. IEEE Conf. Comp. Vision Pattern Recognition, pp.1007–1013, 2009.
[9] J. Matas, O. Chum, M. Urba, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," Proc. British Machine Vision Conf., pp.384–396, 2002.
[10] C. Arth, C. Leistner, and H. Bischof, "Robust local features and their application in self-calibration and object recognition on embedded systems," Proc. IEEE Conf. Comp. Vision Pattern Recognition, vol.1-8, pp.3205–3212, 2007.