Patch-based object tracking using the local robust histogram and background estimation

Ruitao Lu[1, a], Wanying Xu[2,b], Yongbin Zheng[3,c], Shengjian Bai[4,d], Xinsheng Huang[5,e]

[1,2,3,4,5] College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha 410073, China

[a]lrt19880220@163.com, [b]wy.xu@163.com, [c]zhengyongbin@gmail.com, [d]shengjian.bai@gmail.com [e]huangxinsheng@163.com

Keywords: Object tracking; Image segmentation; background estimation.

Abstract. A novel patch-based algorithm for robust object tracking is proposed in this study. The patches of the appearance model are represented by the proposed local robust histogram. Then, the background model is constructed by a set of new spatial probability maps in a surrounding “context window”. For a new testing frame, the vote maps that are obtained by matching the target patches independently are fused for determining the new location of the object. Then, a two-stage estimation method is proposed to estimate the probability of the pixels belonging to the target in the new location. The patches are classified into foreground patches and occluded patches. At last, a dynamic updating scheme is proposed to address appearance variations and alleviate tracking drift. Experiments and evaluations on various challenging image sequences are performed, and the results show that the proposed algorithm performs favorably against other state-of-the-art methods.

Introduction

Object tracking is one of the most important issues in computer vision, which has been widely applied in surveillance, activity analysis, classification and recognition from motion [1]. Although significant progress has been made over the past few decades, object tracking still remains challenging due to significant appearance changes caused by shape deformation, pose variation, motion blur, occlusion and illumination.

In recent years, significant research has been performed regarding patch-based tracking methods [2-7]. These methods perform well in pose changes and occlusion. Adam et al. [2] developed a fragment-based appearance model that considered the target as a collection of image patches; their tracker performed well with pose changes and occlusion. In reference [3], the foreground appearance was modeled by a small number of articulating blocks for addressing pose variations. Jose Bins et al. [4] presented a patch-based tracking method; each patch was tracked individually, and the individual displacement vectors were combined in a robust manner to obtain the accurate tracking results. Liu et al. [5] used a local sparse coding histogram to model the appearance of patches. In reference [6], the appearance of object was modeled by a distribution field that captured the local information of the patches in gray level. In order to address the drifting problem and occlusion, Zhong et al. [7] proposed a robust visual tracking method via patch-based appearance model driven by context-awareness and attentional selection. The local background estimation was used to update the appearance model which provided the robust tracking results.

However, these methods cannot achieve long-term persistent tracking in ever-changing environments. The patches of the appearance model in these methods are represented by histogram or gray vector. If the object is corrupted by appearance changes or occlusion, the similarity of this representation may be significantly affected.

In this study, a patch-based tracking method using the local robust histogram and background estimation is presented. The contributions of this work are threefold: (1) we present a local robust histogram that is constructed by exploiting the spatial relationships to avoid the unstable representation of the original histogram; (2) we propose a two-stage estimation method to estimate the probability belonging to the target of the pixels in the new location. The first stage estimates the location probability of the new location by the Haralick’s cubic facet model [8], and the second stage estimates the gray density probability using the kernel density estimation technique; (3) we propose an effective updating scheme to update the object patches and the background model.

The proposed tracking algorithm

Local Robust Histogram.

The appearance model of object is represented by a collection of image patches in this study, as shown in Fig. 1. To exploit the spatial relationships among the patches, we construct the local robust histogram in this study. The local robust histogram $H_l(F_i)$ for the foreground patch F_i is the intersection of the raw histograms $H_r(F_i^N)$ of patches F_i^N neighboring the patch F_i.

$$\begin{aligned} H_l(F_i(b)) = {}&[\min(H_r(F_i^N(b))), \max(H_r(F_i^N(b))), \\ &median(H_r(F_i^N(b))), mean(H_r(F_i^N(b)))] \end{aligned} \tag{1}$$

where b indexes the bins.

Background model.

Because background information is beneficial to improve tracking stability and accuracy, we construct the background model based on a surrounding “context window” of object, as shown in Fig. 1. The surrounding area of the object is segmented into $N_C \times N_R$ patches with its spatial coordinates. Each patch is represented by the raw histogram with N bins. For each bin, we construct a spatial probability map by resembling the probability of all patches. Then, the background model can be represented by a set of spatial probability maps: $M_p^b(\cdot,\cdot), b = 1,2,...,N$, the size of M_p is $N_C \times N_R \times N$.

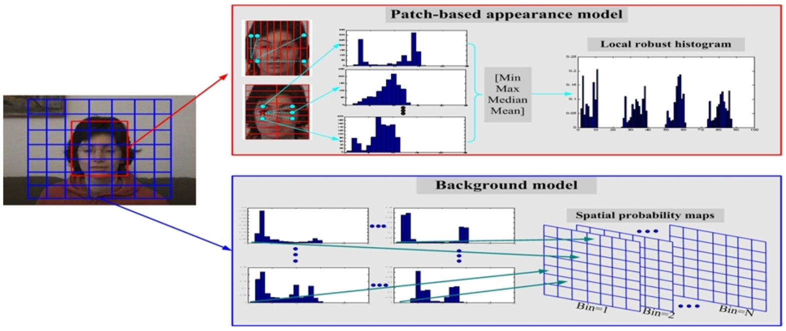


Fig. 1. Illustration of the patch-based appearance model of the object and corresponding local background model.

Matching And Decision Fusion.

For a new testing frame t, we obtain a vote map $V_i^P(\cdot,\cdot)$ for every template patch F_i based on the Chi-square distance. To handle outliers resulting from occluded patches, we obtain the final recognition result by combing a set of vote maps $V^P = \left\{V_i^P(\cdot,\cdot) \mid i = 1,2,...,N_p\right\}$. More specifically, at each location (x, y), we order the corresponding vote values $V^P(x, y) = \left\{V_i^P(x, y) \mid i = 1,2,...,N_p\right\}$, and choose the Q-th smallest values $\left\{V^P(x, y)_r \mid r = 1,2,...,Q\right\}$ to construct the fusion confidence map: $C(x, y) = \sum_{r=1}^{Q} V^P(x, y)_r$. Typically, Q is set as 25%. The robust estimator is designed as follows:

$$(x^*, y^*) = \arg\min_{(x,y)} C(x, y) \tag{2}$$

Two-Stage Background Estimation.

According to the background model, we present a two-stage estimation method to estimate probability of the pixels belonging to the target, as shown in Fig. 2.

The first stage estimates the location probability of the new location by the Haralick's cubic facet model [8]. We use a bivariate cubic function f to approximate the underlying spatial probability maps surface. Let S denotes a symmetric 2D neighborhood defined on $R\times C$, the set of discrete orthogonal polynomials $g_i(r,c), i=1,...,10$ for the cubic function over $R\times C$ is 1, r, c, r^2-2, rc, c^2-2, $r^3-(17/5)r$, $(r^2-2)c$, $r(c^2-2)$, $c^3-(17/5)c$. For every spatial probability map $M_p^b(r,c), b=1,...,N$, the bivariate cubic function $f_b(r,c)$ is described by

$$f_b(r,c)=\sum_{i=1}^{10} K_i^b g_i(r,c), b=1,...,N \tag{3}$$

where $K_i^b, i=1,...,10$ are coefficients for the bivariate cubic function. The coefficients can be computed by

$$K_i^b=\frac{\sum_{(r,c)\in S} g_i(r,c)M_p^b(r,c)}{\sum_{(r,c)\in S} g_i^2(r,c)}, i=1,...,10, b=1,...,N \tag{4}$$

The second stage estimates the gray density probability of the new location using the kernel density estimation technique. For a gray value v in position (r,c) of the new location, its density probability belonging to the background is calculated by

$$f_v(r,c)=\frac{1}{C_h}\sum_{t\in R_v} K_H(\left|\frac{v-v_t}{h}\right|^2) \tag{5}$$

where $K_H(\cdot)$ is the Gaussian kernel function, h is the brand width. R_v is the effective region for the parameter v and C_h is a normalization constant.

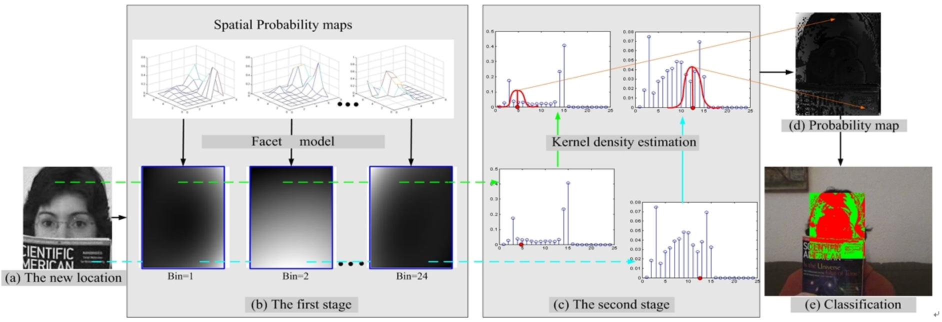


Fig.2. Illustration of the two-stage estimation method. (a) new location of the interest object. (b) the first stage estimates the location probability. (c) the second stage estimates the gray density probability. (d) the probability map of new location.(e) the result of classification.

Updating Scheme.

The pixels of the new location are classified as foreground if they do not adhere to the background model. The classifier can be formulized as follows:

$$Lab(r,c)=\begin{cases} fore\_pixel, & if \ \ f_v(r,c)<q \\ back\_pixel, & otherwise \end{cases} \tag{6}$$

where q is a pre-defined threshold. The patch is considered as the foreground patch if $\frac{sum(fore\_pixel)}{sum(fore\_pixel)+sum(back\_pixel)}>h$. Otherwise, the patch is deemed to occluded patch. In this study, we present a threshold-based updating method to update the foreground patches. For a certain patch, if the number of its neighboring foreground patches is more than 3, the old local robust histogram is replaced with the new one in current frame; if the number of its neighboring foreground patches is less than 2, we assume the patch is occluded and delete the patch from the dictionary template patches.

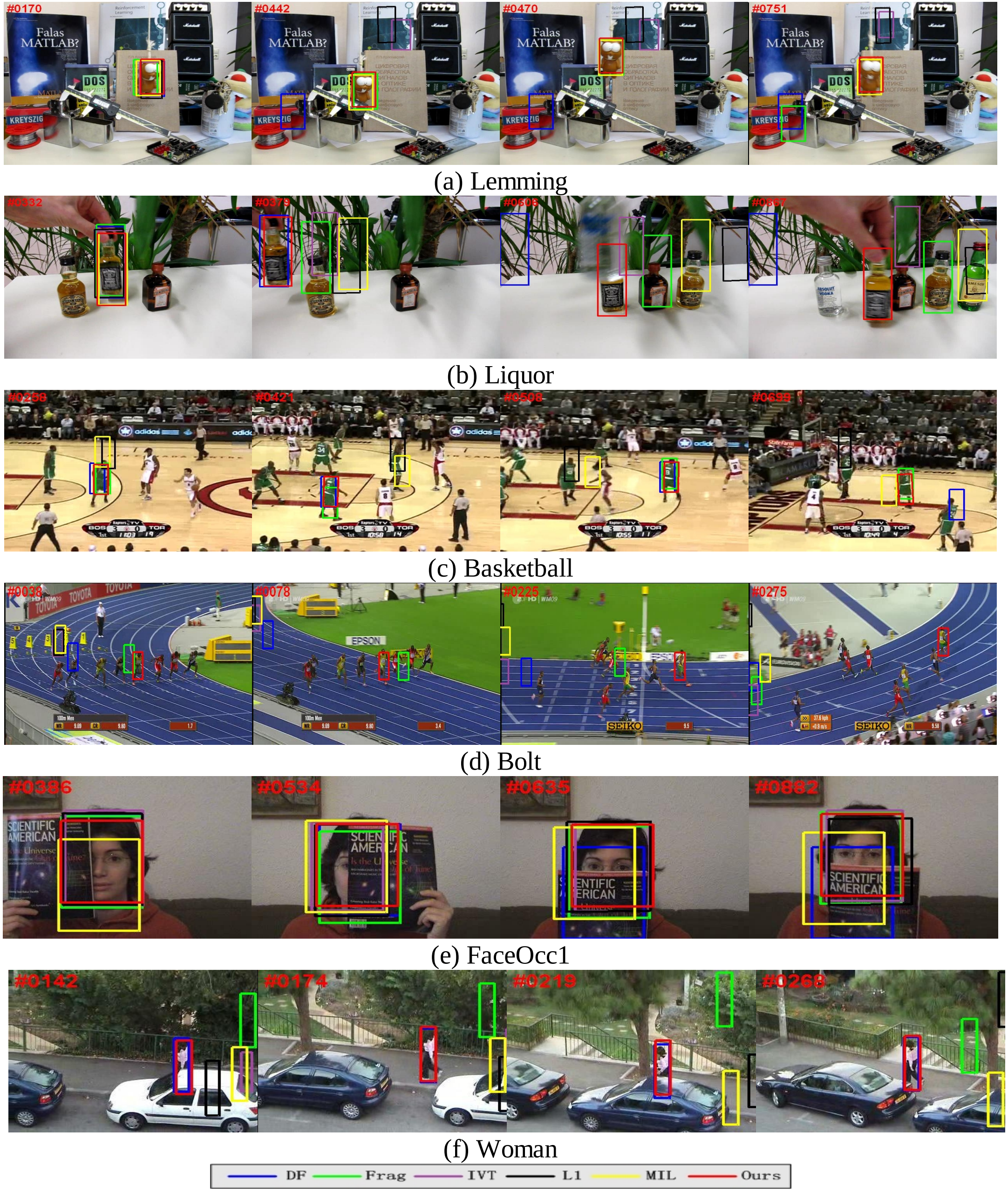
The local background model is updated using a component-wise convex combination of the old background model and new calculated background model:

$$M_p^b(\cdot,\cdot)_N = a M_p^b(\cdot,\cdot)_O + (1-a) M_p^b(\cdot,\cdot)_N , \quad b = 1,2,\ldots,N \tag{8}$$

which a is a forgetting factor.

Experimental Results

Our approach did not use any motion model to predict the target position. The size of search window was defined by enlarging the target region by one third of its size in each direction. The bins N was defined as N=24. The surrounding "context window" was obtained by enlarging the target region by half of its size in each direction. The surrounding area of the object was segmented into 7×7 patches. For comparison, we evaluate our tracker against five state-of-art tracking methods on six challenging sequences including the Frag[2], DFT[6], L1[9], MIL[10] and IVT[11]. Both qualitative and quantitative evaluations are presented in this section.



(a) Lemming

(b) Liquor

(c) Basketball

(d) Bolt

(e) FaceOcc1

(f) Woman

Fig. 3. Qualitative evaluation of six algorithms on six challenging image sequences.

Qualitative Comparison.

Lemming, Liquor: The main challenging factors of the sequences include pose changes and partial occlusion. In the Lemming sequence, L1 and IVT methods drift away from the target (#441, #470). The proposed tracker performs well in this sequence especially when significant pose changes or partial occlusion occurs (#470, #751). In the Liquor sequence, most of the trackers drift away from the target and perform ineffectively in handling objects with large-scale variation (#378, #606). The proposed tracker outperforms other trackers (#606, #867) because the constructed patch-based appearance model is robust to pose variations.

Basketball, Bolt: In the Basketball sequence, IVT, MIL and L1 trackers drift to the background region when the target undergoes pose variations in cluttered background (#258, #421). The patch-based methods perform well in this sequence (#508, #699). In the Bolt sequence, most of the trackers fail to track the target when similar objects appear around the target or significant appearance changes occurs (#38 #78, #225). Only the proposed tracker tracks the target object stably throughout the entire sequence.

Woman, FaceOcc1: A face in the FaceOcc1 is heavily occluded by a magazine. Because there is no drift alleviation scheme due to its holistic sparse representation, L1 tracker can not track effectively in this sequence (#882). The proposed tracker achieves the lower drifting errors than Frag tracker. In the Woman sequence, the target undergoes long-term partial occlusion (#142, #219, #248) and pose changes (#174). Only DF and the proposed algorithm can track the target in this sequence. The two-stage estimation method provides supervision to an analysis of possible occlusion of the object.

Quantitative Comparison.

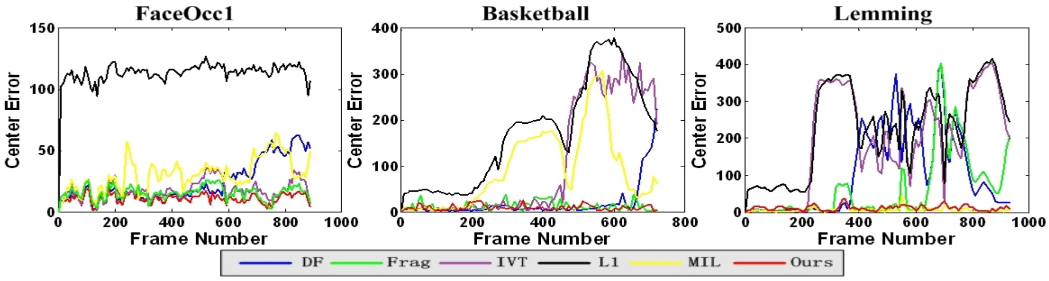


Fig. 4. Center location error evaluation of six algorithms on some image sequences.

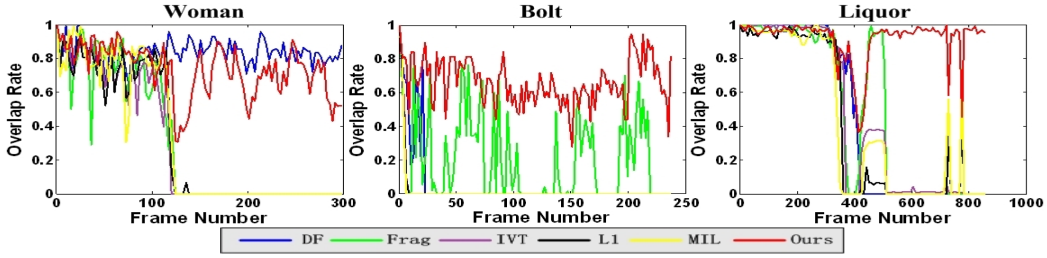


Fig. 5.Overlap rate evaluation of six algorithms on some image sequences.

For quantitative evaluation, two evaluation criteria are used to assess the performance of tracking algorithms. The average errors are presented in Table 1. The proposed tracker achieves the lower drifting errors than the others tracking algorithms in almost all consequences. Fig. 4 shows the center location errors in pixels of each algorithm for some image sequences.

Table 1. the Overlap rate (ORE). the **Bold** fonts indicate the best performance.

Sequence	IVT	L1T	DF	Frag	MIL	Ours
Basketball	0.10	0.23	0.61	0.62	0.22	**0.67**
Bolt	0.02	0.01	0.05	0.20	0.02	**0.66**
Woman	0.27	0.28	**0.83**	0.27	0.29	0.70
FaceOcc1	0.73	0.73	0.69	0.81	0.60	**0.86**
Lemming	0.20	0.19	0.33	0.44	**0.71**	0.66
Liquor	0.45	0.40	0.45	0.49	0.42	**0.81**

In addition, the overlap rate [12] is used to evaluate the stability of each algorithm. The overlap rates of each tracking algorithm for some image sequences are illustrated in Figure 5. The success of our tracker can be attributed to the effective patch-based appearance model with spatial representation and the background estimation provides effective mechanism to update the dictionary template patches.

Conclusions

A new patch-based approach for object tracking is presented in this study. The representation of the proposed local robust histogram is more robust to occlusion and appearance changes. The two-stage estimation method based on the background model provides supervision to an analysis of possible occlusion of the object. Both quantitative and qualitative evaluations on challenging image sequences against several state-of-the-art algorithms demonstrate the accuracy and the robustness of the proposed tracker.

Acknowledgements

This work was financially supported by the National Science Foundation of China (No. 61403412)

References

[1] LI X, HU W, SHEN C, ZHANG Z. A Survey of Appearance Models in Visual Object Tracking. ACM Transactions on Intelligent Systems and Technology, 2013, 4(4).

[2] ADAM A, RIVLIN E, and SHIMSHONI I. Robust fragments-based tracking using the integral histogram. In Proc. CVPR, 2006: 798–805.

[3] NEJHUM S, HO J, YANG M. Online visual tracking with histograms and articulating blocks. Computer Vision and Image Understanding, 2010: 901–914.

[4] JOSE B, LEANDRO L, CLAUDIO R. Target Tracking Using Multiple Patches and Weighted Vector Median Filters. J Math Imaging Vis, 2013,45: 293–307.

[5] LIU B, HUANG J, KULIKOWSKI C, YANG L. Robust tracking using local sparse appearance model and K-selection. In Proc. CVPR, 2011: 1–8.

[6] LAURA S, ERIK L. Distribution Fields for Tracking. In Proc. CVPR, 2012.

[7] ZHONG B, CHEN Y, SHEN Y, et al. Robust tracking via patch-based appearance model and local background estimation. Neurocomputing, 2014,123: 344–353.

[8] HARALICK R. Digital step edges from zero crossing of second directional derivatives. IEEE Trans. Pattern Anal. Mach. Intell., 1984,6(1): 58–68.

[9] MEI X, and LING H. Robust visual tracking using 1 minimization. In Proc. ICCV, 2009: 1436–1443.

[10] BABENKO B, YANG M.-H, and BELONGIE S. Visual tracking with online multiple instance learning. In Proc. CVPR, 2009: 983–990.

[11] ROSS D, LIM J, LIN R, and YANG M. Incremental learning for robust visual tracking. Int. J. Comput. Vis., 2008, 77: 125–141.

[12] EVERINGHAM M, VAN L, WILLIAMS C-K, WINN J, and ZISSERMAN A. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis., 2010, 88(2): 303–338.