Robust Object Tracking using Probabilistic Hypergraph Ranking and Superpixels

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: Visual tracking, transductive learning, hypergraph ranking, superpixel.

Abstract. Online object tracking is a challenging issue because the appearance of an object tends to change due to intrinsic or extrinsic factors. In this study, we propose a robust tracking algorithm based on probabilistic hypergraph ranking and superpixels. The probabilistic hypergraph is constructed by mid-level visual cues and their spatial relationships. Then, the confidence map at mid-level cues is obtained by hypergraph ranking analysis, which takes the high order intrinsic relationships of superpixels into account. Third, Object tracking is formulated as a transductive learning issue, and the optimal target location is determined by maximum a posterior estimation on the ranking scores. Finally, a dynamic updating scheme is proposed to address appearance variations and alleviate tracking drift. A series of experiments and evaluations on various challenging sequences are performed, and the results show that the proposed algorithm performs favorably against other existing 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, classification, activity analysis and recognition. Typically, a visual tracking system consists of four modules [1]: object initialization, appearance model, motion model, and object localization. In recent years, significant research has been performed regarding discriminative tracking methods [2-8]. These methods consider visual object tracking as a binary classification problem by identifying targets from backgrounds. Both classic and recent machine-learning algorithms are used to promote the performance of these tracking methods [2-8]. These methods generally assume that backgrounds and targets are separated linearly; however, this assumption is violated when the object undergoes significant changes in a real application. The classifier is constructed only by a few expensive labeled samples, and a large amount of unlabeled samples are abandoned.

Graph-based transductive learning methods [9-14] study the intrinsic geometric structure of both labeled and unlabeled samples and can thus explore the affinity relationships among vertices. Zhang et al. [13] proposed a graph-based learning method for tracking in which a graph structure is designed to reflect the properties of the sample distributions. In [14], a graph-based transductive learning method was proposed in both variable indoor and outdoor scenes. However, these methods only consider the pairwise interactions between vertices [10,12]. In addition to the relationships of two individual vertices, their corresponding contexts, which contain local information, should be considered as well.

Motivated by above-mentioned discussions, we present a robust tracking algorithm based on probabilistic hypergraph ranking and superpixels. We introduce hypergraph modeling into the object-tracking process for the first time. The probabilistic hypergraph is constructed by encoding the local affinity information of superpixels. Then, we formulate the object tracking problem as a transductive learning issue, and the optimal target location can be obtained by maximum a posterior estimation on the ranking scores. Finally, a dynamic updating scheme is proposed to address appearance variations and alleviate tracking drift. We conduct numerous experiments on challenge sequences to demonstrate the effectiveness of our proposed method.

Confidence map based on Probabilistic hypergraph ranking

Hypergraph ranking [12] is a universal ranking algorithm used to rank samples along their high order intrinsic manifold structure. In this section, the confidence map at mid-level cues is computed by the hypergraph ranking to construct our discriminative appearance model.

Probabilistic hypergraph ranking.

Probabilistic hypergraph $G=(V,E,w)$ is formed by a vertex set V, hyperedges E which denotes a family of subsets e, and a positive hyperedge weight $w(e)$. The incidence matrix H of probabilistic hypergraph can be represented by: $h(v_i,e_j)=A(i,j)$, if $v_i \in e_j$, and $h(v_i,e_j)=0$, otherwise, where $A(i,j)$ is a kernel function used for measuring the similarity between v_i and the 'centroid' vertex of the hyperedge e_j. The hypergraph weight $w(e)$ can be defined as $w(e_i)=\sum_{v_j \in e_i} A(i,j)$.

The degree of each vertex can be defined $d(v)=\sum_{e\in E} w(e)h(v,e)$. For a hyperedge $e \in E$, its degree is $d(e)=\sum_{v\in V} h(v,e)$. We use D_v, D_e and W to denote diagonal matrices of the vertex degrees, the hyperedge degrees and the hyperedge weights, respectively. A hypergraph is constructed as show in Fig.2(b).

To improve the effect of the feedback information and introduce the diagonal constraint, we defined the cost function $\Omega(f)$ as:

$$\Omega(f)=\frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)h(u,e)h(v,e)}{d(e)}\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2+u\sum_{v\in V}\|f-y\|^2 \quad (1)$$

where the vector y is the indication vector , the f is the assigned sample value to be learned, and u is a tradeoff parameter. Let $\Theta=D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}$, the result ranking function can be written as

$$f=(1-g)(I-g\Theta)^{-1}y \quad (2)$$

where $g=1/(1+u)$.

Confidence map based on superpixels.

In our framework, surrounding area of the target is segmented into lots of superpixels [15] for representing the vertices. A Location-Adjacent Hypergraph $G_l=(V,E_l,w_l)$ is constructed in this paper. The location-adjacent hyperedge e_l is composed of a 'centroid' superpixel and neighboring superpixels in image coordinate system. The vertices on the four sides of the surrounding area are considered to be adjacent.

We use the Mean Shift algorithm [16] to generate the indicator vector y. The M frames with the ground truth are collected for training. The surrounding area of the target is segmented into N_t superpixels such as $\mathbf{a}_{t,r}(t=1,...,M,r=1,...,N_t)$. A feature pool $F=\{\mathbf{a}_{t,r}\,|\,t=1,...,M,r=1,...,N_t\}$ is cluster into $Cluter_i(i=1,...,n)$ in feature space. The positive cluster $Cluter^+{}_j(j=1,...,N_+)$ is defined as $S_j^+/S_j^-<l\,(l>1)$, where S^+ denotes that $Cluter_i$ contains the local areas inside the target area, and S^- represents the local areas outside the target area. Analogously, the negative cluster is $Cluter_j^-(j=1,...,N_-)$ if $S_j^+/S_j^-<l$. With a new test M_t -th frame, the positive query set Q^+ is obtained by k-NN classifier (k=1) based on the positive cluster:

$$Q^+=\{Q_j^+\,|\,Q_j^+=\arg\min_{r=1,...,N_t}(dis(\mathbf{a}_{M_t,r},Cluter_j^+)),\ j=1,...,N_+\}. \quad (3)$$

The negative query set Q^- contains two parts: $Q^-=Q_1^-\,\mathbf{U}\,Q_2^-$. Q_1^- is obtained from negative cluster, and Q_2^- is obtained from the boundary of surrounding area:

$$\begin{cases} Q_1^- = \{Q_{1,j}^- \mid Q_{1,j}^- = \arg\min_{r=1,\ldots,N_t}(dis(\mathbf{a}_{M_t,r}, Cluter_j^-)),\ j = 1,\ldots,N_- \} \\ Q_2^- = \{\mathbf{a}_{M_t,r} \mid \mathbf{a}_{M_t,r} \in boundary, r = 1,\ldots,N_t\} \end{cases} \tag{4}$$

The indication vector y can be defined as $y_i = 1/|Q^+|$, if $v_i \in Q^+$, $y_i = -1/|Q^-|$, if $v_i \in Q^-$, and $y_i = 0$ otherwise. The confidence map at superpixels can be computed by (2) (see Fig.1).

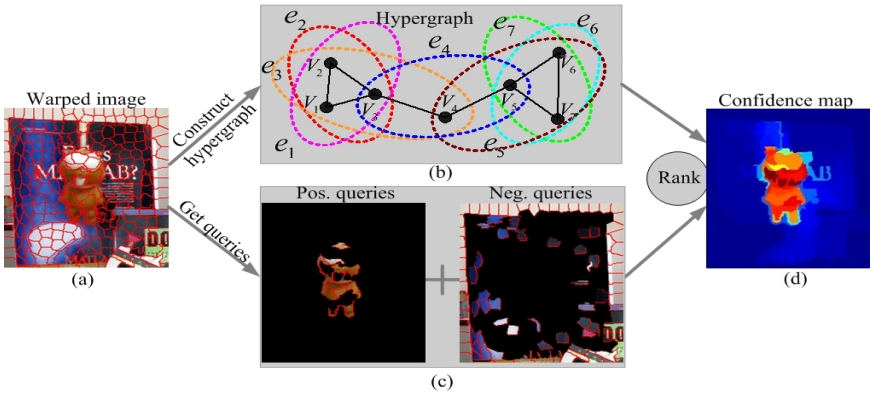

Fig. 1 Process of creating confidence map. (a) warped image in a new frame; (b) the constructed probabilistic hypergraph based on superpixels; (c) positive query set and negative query set; (d) the final confidence map computed by hypergraph ranking.

Proposed tracking Algorithm.

The visual track problem can be considered as a Bayesian inference task in a Markov model [5,14,17,18] with hidden state variables. Given a set of observed images $\mathbf{Y}_t = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_{t-1}, \mathbf{y}_t\}$ at the t-th frame, the hidden state $\mathbf{x}_t$ can be estimated:

$$p(\mathbf{x}_t \mid \mathbf{Y}_t) \propto p(\mathbf{y}_t \mid \mathbf{x}_t) \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid \mathbf{Y}_{t-1}) d\mathbf{x}_{t-1} \tag{5}$$

where $p(\mathbf{y}_t \mid \mathbf{x}_t)$ is the observation model and the $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ represents the dynamic model.

We model the dynamic model between two consecutive frames with affine transformation. Let $\mathbf{x}_t = \{x_t, y_t, q_t, s_t, a_t, f_t\}$, where $x_t, y_t, q_t, s_t, a_t, f_t$ denote x, y translation, rotation angle, scale, aspect ratio, and skew direction respectively. We use a random walk model for the state transition, i.e., $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = N(\mathbf{x}_t; \mathbf{x}_{t-1}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal covariance matrix.

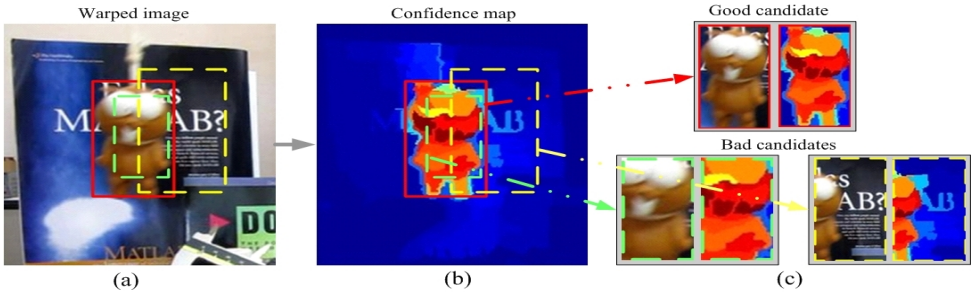

Fig. 2 Proposed observation model. (a) warped image; (b) final confidence map; (c) the good target candidate is shown in red rectangle and others are bad target candidates.

In confidence map, a good target candidate not only has a higher confidence value, but also covers more parts of foreground regions. The observation model (see Fig.2) is defined as:

$$p(\mathbf{y}_t^i \mid \mathbf{x}_t^i) = \sum_{(j,k) \in M_i} f(j,k) \times (|\mathbf{s}_t^i| / |\mathbf{s}_{t-1}|) \tag{6}$$

where $|\mathbf{s}_t^i|$ represents the real area size of target state $\mathbf{x}_t^i$. The optical target state $\hat{\mathbf{x}}_t$ can be obtained by Maximum a Posteriori (MAP) estimation over samples:

$$\hat{\mathbf{x}}_t = \arg\max_{x_t^i} p(\mathbf{y}_t^i \mid \mathbf{x}_t^i)\, p(\mathbf{x}_t^i \mid \mathbf{x}_{t-1}),\ i = 1,2,3,...,N \tag{7}$$

where $\mathbf{x}_t^i$ is the i-th sample of the state $\mathbf{x}_t$. For every U frames, a new frame is put into the training dataset and the oldest one is deleted. And we update the appearance model for clustering the feature pool for every W frames.

Experimental Results

Our approach is implemented in MATLAB 2011 on a Core 2.0GHz Dual Core PC with 2 GB memory. Each image observation is normalized to 32×32 pixels, and the number of samples is set to 300. The surrounding region is set 1.5 times of the size of target area. The frames of training and the number of superpixels are set to 5 and 300. The threshold l is set to the range of 2 and 4, and the parameter g is set to 0.1. The update frequency W is 8, and the spacing interval U is 4. For comparison, we evaluate our tracker against seven state-of-art tracking methods on eight challenging sequences including the IVT[17], L1[18], MIL[6], Struck[7], OAB[2], SemiT[3] and PN algorithms [8]. Both qualitative and quantitative evaluations are presented in this section.

Qualitative Comparison.

pose

The Lemming sequence as shown in Fig. 3(a), the proposed method, MIL and OAB perform better than the other methods. OAB and MIL methods work well as they select the most discriminative Haar-like features for object representation which can well handle pose variation and shape deformation. In the Bike sequence as shown in Fig. 3(b), IVT, L1 and SemiT perform poorly as the background has a seriously interfere. Compared to Sturck and OAB, our method distinguishes target parts from background blocks more precisely (#205).

Occlusion

Fig. 3(c) presents the tracking results on the Woman sequence with long-term partial occlusion (#0132, #0229). MIL tracker does not perform well as the generalized Haar-like features used in the tracker are less effective to occlusion. The proposed tracker and Struck tracker perform better than the other methods. Some tracking results on the Liquor sequence are shown in Fig. 3(d). Most methods fail to track the target well when the target undergoes occlusion (#606, #776). As our hypergraph-based discriminative appearance model learns the appearance model of both superpixels and their manifold structure, it is able to detect the target all heavy occlusions.

Illumination change and scale change

Figure 3 (e) presents the tracking results in the Singer1 sequence. The proposed method and TLD perform better than the other methods. OAB, SemiT and Struck trackers don't adapt to scale change (#288), and L1 tracker achieves a high tracking error . Figure 3(f) presents the tracking results in the Carscale sequence with large scale variation. Most trackers drift away from the target when the objects with large scale variation. Our tracker outperforms other tracker as the proposed hypergraph-based discriminative appearance model making use of mid-level visual cues and their high order intrinsic structure.

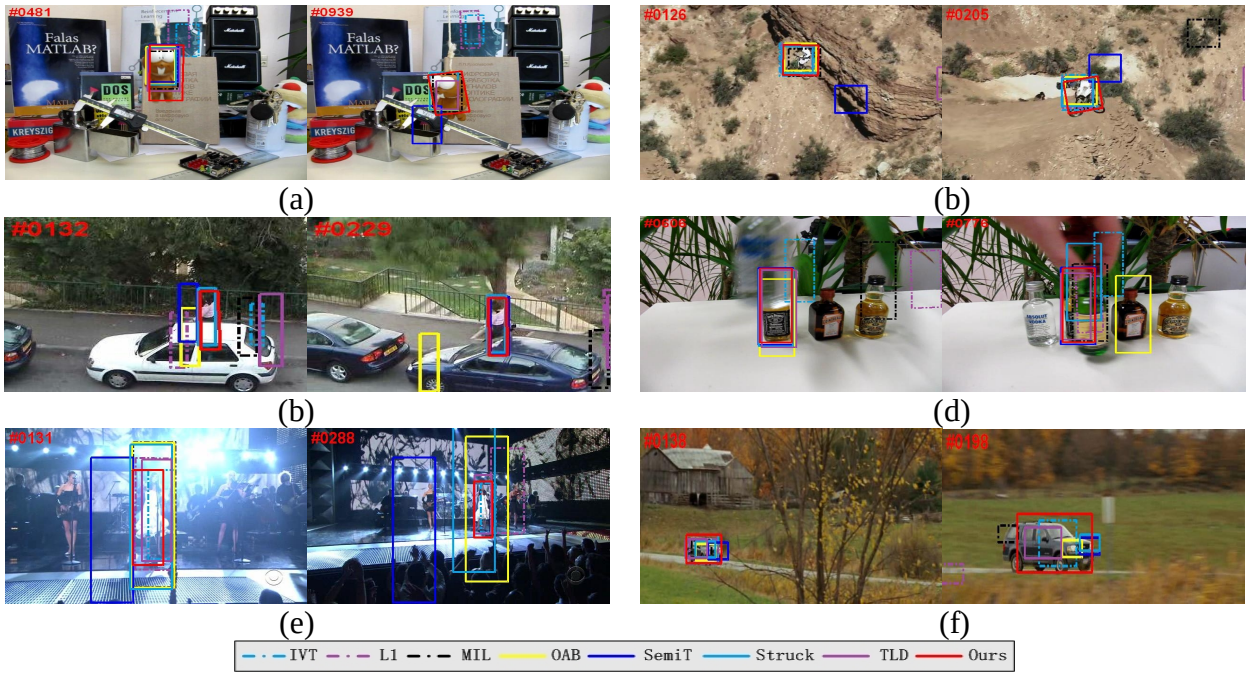

Fig. 3. Qualitative evaluation of seven algorithms on eight challenging image sequences.

Qualitative Comparison.

For quantitative evaluation, we employ two evaluation criteria to assess the performance of tracking algorithms.. The average center location errors are presented in Table 1. Our tracker achieves the lower drifting errors than the others tracking algorithms in almost all consequences. In addition, we employ the overlap rate [19] to evaluate the stability of each algorithm. Table 2 presents the average overlap rates. The success of our tracker can be attributed to the effective discriminative appearance model with mid-level visual representation.

Table 1 Center location error (CLE) (in pixels). **Bold** fonts indicate the best performance.

Image sequence	IVT	L1	MIL	Struck	OAB	SmeiT	TLD	Ours
Lemming	181	214	12.1	37.8	18.1	161	16.0	**9.0**
Singer1	11.7	146	16.4	14.5	12.9	98.5	7.9	**5.6**
Carscale	11.6	106	33.2	35.7	30.0	26.8	21.6	**6.2**
Woman	163	124	116	**3.6**	34.7	19.5	130	9.5
Liquor	118	244	141	91.0	68.6	64.2	37.6	**15.6**
Bike	7.4	50.9	73.0	8.6	12.0	56.1	216	**5.2**

Table 2 Overlap rate (ORE). **Bold** fonts indicate the best performance

Image sequence	IVT	L1	MIL	Struck	OAB	SmeiT	TLD	Ours
Lemming	0.14	0.13	0.65	0.48	0.60	0.14	0.44	**0.71**
Singer1	0.57	0.29	0.36	0.37	0.34	0.17	**0.73**	0.72
Carscale	0.65	0.47	0.41	0.41	0.40	0.43	0.46	**0.79**
Woman	0.18	0.18	0.19	**0.76**	0.48	0.39	0.16	0.69
Liquor	0.23	0.20	0.22	0.42	0.45	0.51	0.52	**0.76**
Bike	0.74	0.73	0.46	0.71	0.64	0.45	0.71	**0.79**

Conclusion

This paper presents a novel tracking algorithm based on probabilistic hypergraph ranking and superpixels. The object tracking is formulated as a transductive learning problem and the most target location is obtained by MAP estimation based on the confidence map. Both quantitative and qualitative

evaluations 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] X Li, W Hu, C Shen, Z Zhang. "A Survey of Appearance Models in Visual Object Tracking," ACM Transactions on Intelligent Systems and Technology, Vol. 4, No. 4, Article 58, Sep. 2013.
[2] H. Grabner, M. Grabner, and H. Bischof. Real-Time Tracking via On-line Boosting. In Proc. BMVC, 2006, pp. 47-56.
[3] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in Proc. ECCV, 2008, pp. 234–247.
[4] S. Avidan, "Support vector tracking," IEEE TPAMI, vol. 26, no. 8, pp. 1064–1072, 2004.
[5] K. Zhang, L. Zhang, and M.-H. Yang. Real-time Compressive Tracking. in Proc. ECCV, 2012, pp. 866-879.
[6] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in Proc. CVPR, 2009, pp. 983–990.
[7] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in Proc. ICCV, 2011, pp. 263–270.
[8] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in Proc. CVPR, 2010, pp. 49–56.
[9] D. Zhou, J. Weston, and A. Gretton, "Ranking on data manifold," in Proc. NIPS, 2004, pp.169–176.
[10] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M. H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (IEEE, 2013), pp. 3166-3173.
[11] Xi Li, W Hu, C Shen, A Dick, Z Zhang, "Context-Aware Hypergraph Construction for Robust Spectral Clustering," IEEE Transactions on knowledge and data engineering.
[12] Q Liu , Y Huang, D N.Metaxas. "Hypergraph with sampling for image retrieval," Pattern Recognition, 44, pp. 2255–2262, 2011.
[13] X. Q. Zhang, S. Y. Chen, "Graph-Embedding-Based Learning for Robust Object Tracking," IEEE Transactions on Industrial Electronics, 61, pp. 1072-1084, 2014.
[14] Y. F. Zha, Y. Yang, and D. Y. Bi, "Graph-based transductive learning for robust visual tracking," Pattern Recognition, 43, pp: 187–196, 2010.
[15] Achanta, R., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S.: 'Slic superpixels'. Technical report, EPFL, Tech.Rep. 149300, 2010. 3
[16] D. Comaniciu, V. Ramesh, and P Meer, "Real-Time tracking of Non-Rifgid Objects using MeanShift," in Proc. CVPR, 2000, pp. 142–149
[17] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," Int. J. Comput. Vis., vol. 77, nos. 1–3, pp. 125–141, 2008.
[18] X. Mei and H. Ling, "Robust visual tracking using 1 minimization," in Proc. ICCV, 2009, pp. 1436–1443.
[19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.