Part Context Learning for Visual Tracking

Guibo Zhu
gbzhu@nlpr.ia.ac.cn
Jinqiao Wang
jqwang@nlpr.ia.ac.cn
Chaoyang Zhao
chaoyang.zhao@nlpr.ia.ac.cn
Hanqing Lu
luhq@nlpr.ia.ac.cn

National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.

Context information is widely used in computer vision for tracking arbitrary objects[1, 3, 4]. Global context cannot deal with the object deformation problem, while the local part context interactions are relatively stable. When the target appearance changes gradually, the intrinsic property of internal interaction between the parts inside object and context interaction between object and background are relatively stable in spatio-temporal 3D space of tracking.

To explore the structure property and stable relationship for overcoming complex environments, we propose a novel part context tracker. The Part Context Tracker (PCT) consists of an appearance model, an internal relation model and an context relation model. The internal relation model formulates the temporal relations of the object itself or the in-object parts themselves and the spatio-temporal relations between the object and in-object parts. The context relation model constructs the spatio-temporal relations between the in-object parts and the context parts and the temporal relations of the context parts themselves. Hence the physical properties and the appearance information are considered in the optimization process through parts and relations. The contributions are as follows:

(1) We first propose a unified context framework which formulates the single object tracking as a part context learning problem.

(2) The in-object parts and context parts are selected so that we not only pay attention to the appearance of object, but also focus on the relations among the object, the in-object parts and the context parts.

(3) A simple yet robust update strategy using median filter is utilized, thereby enabling the tracker to deal with appearance change effectively and alleviate the drift problem.

Our framework not only models the object with in-object parts, but also incorporates the interaction between the object and background with context parts. The deformable configuration [2, 5] together with the temporal structure of these parts are also considered in.

In Fig. 1, with the object bounding box as the root R, the in-object parts I are defined as the parts selected inside R, which covers part of the object appearance. The context parts C are selected from the overlapping area between the object and the background. For a target with K in-object parts and M context parts, the configuration is denoted as $B = (B_0, B_1 ... B_K, B_{K+1}, ..., B_{K+M})$. Where B_0 stands for the target bounding box R, $(B_1, .., B_K) \in I$ are the K in-object part boxes, and $(B_{K+1}, ..., B_{K+M}) \in C$ are the M context part boxes. The corresponding features of the root and parts are represented as $X = (x_0, ..., x_K, x_{K+1}, ..., x_{K+M})$. In a word, our framework models the object with three components:

$$M = M_A + M_I + M_C, \tag{1}$$

where M_A, M_I and M_C are the appearance model, the internal relation model and the context relation model respectively.

For online tracking, an appearance model is essential. It represents the intrinsic property of one object or the discriminative information between the object and background. To better mine the information, we factorize the appearance model M_A as Eq. (1):

$$M_A = A_R + A_I + A_C \tag{2}$$

where A_R, A_I and A_C are the global root appearance model, in-object parts appearance model and context parts model separately.

In addition to the appearance model, all spatio-temporal relative stable relations between the object and its corresponding parts frame-to-frame should be utilized in tracking. Therefore we design the internal relation model to formulate the interactions between root and the in-object parts, which includes the spatial constrains and the temporal constrains between them, we define M_I as:

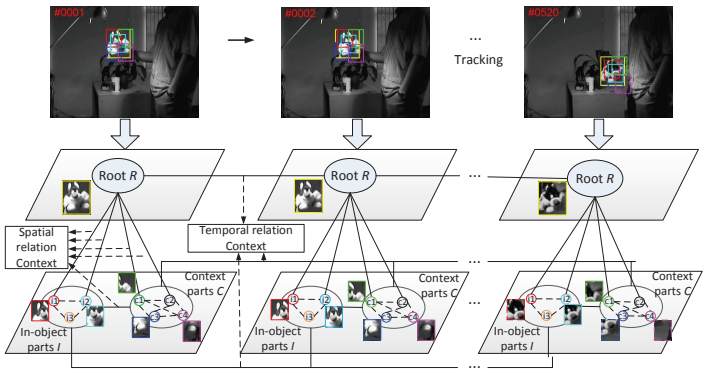

Figure 1: Illustration of our Part Context tracking framework using the "sylvester" video.

$$M_I = S_I + E_R + E_I \tag{3}$$

where S_I, E_R, and E_I are spatial relation between root and in-object parts, temporal relation between root and their historical root, and temporal relation between in-object parts and their historical information respectively.

Except internal relations inside the object, some information in latent intersection area between the object and background is neglected by previous works, such as the partial contour and the object are consensus in motion. To make full use of the information, we formulate the context relation model to express the interactions between root and the context parts, which also includes the spatial and temporal constrains between them. Similar to Eq. (3), we describe the context relation model mathematically as:

$$M_C = S_C + S_{C,I} + E_C \tag{4}$$

where S_C, $S_{C,I}$ and E_C denote spatial relation between root and context parts, spatial relation between in-object parts and context parts, and temporal relation between context parts and their historical information.

Implementation of this method by model definition is described in the paper, as are the details of the model optimization in inference and learning. Our conclusion is that one tracker consists of an appearance model, an internal relation model and an context relation model in a maximum margin structured learning framework, which is robust to certain conditions of occlusion, illumination and out-of-view.

[1] T.B. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, pages 1177–1184. IEEE, 2011.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[3] L. Wen, Z. Cai, Z. Lei, D. Yi, and S.Z. Li. Online spatio-temporal structural context learning for visual tracking. In *ECCV*, pages 716–729. Springer, 2012.

[4] K. Zhang, L. Zhang, M. H. Yang, and D. Zhang. Fast tracking via spatio-temporal context learning. *arXiv preprint arXiv:1311.1939*, 2013.

[5] L. Zhang and L. van der Maaten. Preserving structure in model-free tracking. *IEEE-TPAMI*, 36(4):756–769, 2014.